1-1-1991

Evaluation of IRT anchor test designs in test translation studies.

John Bollwark
University of Massachusetts Amherst

EVALUATION OF IRT ANCHOR TEST DESIGNS IN TEST TRANSLATION STUDIES

A Dissertation Presented

By

John Bollwark

Submitted to the Graduate School
of the University of Massachusetts
in partial fulfillment of the requirements
for the degree of

DOCTOR OF EDUCATION

September 1991

School of Education

EVALUATION OF ANCHOR TEST
DESIGNS IN TEST TRANSLATION STUDIES

A Dissertation Presented

By

JOHN BOLLWARK

Approved as to style and content by:

Ronald K. Hambleton, Chair

Hariharan Swaminathan, Member

Arnold D. Well, Member

Bailey Jackson, Dean
School of Education

ACKNOWLEDGMENTS

I would like to give my deepest thanks to a number of individuals who made my experiences at the University of Massachusetts so rewarding and enjoyable. I first want to thank Professors Ronald Hambleton and Hariharan Swaminathan. Ron took a personal interest in all aspects of my academic career. He provided me with many rewarding professional experiences and continually spurred my interest in the measurement field. His dedication to his students and teaching was greatly appreciated. I especially want to thank Ron for his help and thoughtful comments on this thesis. Hariharan, in addition to providing excellent instruction in statistics, made himself available for constant inquiry and was very supportive of my personal goals. Both Ron and Hariharan have set professional standards that I will continually strive to meet.

In addition, special thanks go to the following individuals. Professor Stan Scarpati for serving as a member of my comprehensive examination committee and for his guidance, support and friendship. Professor Arnold D. Well for taking time from his busy schedule to serve on my thesis committee. Dr. H. Jane Rogers, for her help with computer programming and for just being there as a friend. Also, Peg Louraine, for her help in preparing my comprehensive examination papers and this thesis.

Finally, I would like to thank my wife, Amy. She continually believed in me and provided unconditional support throughout my academic career. Her belief in the importance of education was a source of inspiration on those sometimes difficult days. I will always appreciate the sacrifices she made to allow me to reach my goals.

ABSTRACT

EVALUATION OF IRT ANCHOR TEST
DESIGNS IN TEST TRANSLATION STUDIES

SEPTEMBER 1991

JOHN BOLLWARK, B. S., UNIVERSITY OF MASSACHUSETTS

Ed.D., UNIVERSITY OF MASSACHUSETTS

Directed by: Professor Ronald K. Hambleton

Translating measurement instruments from one language to another is a common way of adapting them for use in a population other than those for which the instruments were designed. This technique is particularly useful in helping to (1) understand the similarities and differences that exist between populations and (2) provide unbiased testing opportunities across different segments of a single population. To help insure that a translated instrument is valid for these purposes, it is essential that the equivalence of the original and translated instrument be established. One focus of this thesis was to provide a review of the history, problems and techniques associated with establishing the translation equivalence of measurement instruments. In addition, this review provided support for the use of item response theory (IRT) in translation equivalence studies. The second and main focus of this thesis was to investigate anchor test designs when using IRT in translation equivalence studies. Simulated data were used to determine the anchor test length required to provide adequate scaling results under conditions similar to those that are likely to be found in a translation equivalence study. These conditions included (1) relatively small samples and (2) examinee ability distribution overlaps that are

more representative of vertical rather than horizontal scaling situations. The effects of these two variables on the anchor test design required to provide adequate scaling results were also investigated.

The main conclusions from this research concerning the scaling of IRT ability and item parameters are: (1) larger examinee samples with larger ability overlaps should be used whenever possible, (2) under ideal scaling conditions of larger examinee samples with larger ability overlaps, relatively good scaling results can be obtained with anchor tests consisting of as few as 5 items (although the use of such short anchor tests is not recommended), and (3) anchor test lengths of at least 10 items should provide adequate scaling results, but longer anchor tests, consisting of well-translated items, should be used if possible.

Finally, suggestions for further research on establishing translation equivalence were provided.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND STATEMENT OF THE PROBLEM

1.1 Background

Adapting tests for use in populations other than those the tests were designed for has its roots in the beginnings of intelligence testing. Psychologists readily saw the potential of intelligence tests for diagnostic and selection purposes, and adapted them from the population for which they were developed for use in different populations of interest. In these early test adaptions, the adaption process usually consisted of a direct translation of a test from one language to another.

More recently, adapting tests for use in populations other than those the test was designed for has been fueled by an interest in providing a basis for cross-population comparisons. Researchers interested in quantifying differences in intelligence and other traits in different populations must rely on test adaptions. Also, in countries such as the United States, issues of test bias have initiated an interest in adapting tests so that they are more relevant and thus "fair" to specific segments of a particular population. The adaption process in these cases should ideally consist of translating a test from one language to another with consideration given to the linguistic and cultural relevance of the translated version and to the "equivalence" of the different versions of the test.

Although validly translating a test from one language to another and establishing the equivalence of the original and translated versions

is a complex process, it is important that the process be better understood since test translations will play an increasingly important role in future testing activities. The main reason for this is that we are increasingly viewing our world from a multicultural perspective and therefore there is a need to (1) understand the similarities and differences that exist between populations and (2) provide unbiased testing opportunities across different segments of a single population. Testing across populations provides a means for accomplishing these goals.

For example, in 1988, the International Assessment of Educational Progress (IAEP) was implemented (Lapointe, Mead, & Phillips, 1989). The goal of this project was to assess achievement in a common core of science and mathematics for 13-year-olds in five countries and four Canadian provinces. In order to accomplish this goal, test items in English were translated into several different languages. Also administered were questionnaires regarding students' school experiences and attitudes towards mathematics and science.

This expensive and time-consuming assessment project was undertaken because of a view that the results would provide insights into differences among populations that influence the attainment of successful educational goals. One result from this study was that students from the United States scored lowest in mathematics achievement while Korean students scored highest. What reason or reasons are responsible for these differences? An answer to this question may be of substantial use in improving mathematics education in the United States and therefore is of vital importance to our society. Without cross-cultural assessment projects such as the IAEP, answers to these types of

questions cannot be obtained. Without a proven methodology for evaluating the equivalence of the original and translated assessment instruments, a valid basis for these types of comparisons remains in question.

1.2 Statement of the Problem

Translating a test from one language to another is a common procedure for adapting a test for use in populations other than those for which the test was designed (Samuda, 1975; Fouad & Hansen, 1987). A test may be translated in order to (1) economically develop a test for use in a population or (2) provide a basis for comparisons between populations or segments of a population. When economically developing a test for use in a population, the original and translated test need not be "equivalent" unless the test scores derived from the translated test are to be referenced in some way back to the source population. In contrast, cross-population comparisons require that at least a portion of the original and translated test items must be "equivalent" in order to make valid comparisons.

A number of different methods for establishing the translation equivalence of test items have been suggested. The different methods arise from differences in (1) the examinee samples used (bilinguals, source language monolinguals, target language monolinguals), (2) the version or versions of a test upon which translation equivalence will be based (source, target, or back-translated version), and (3) whether judgmental or statistical procedures are used.

One of the more promising statistical techniques for establishing translation equivalence is item response models. When an item response model fits the test data from examinee samples in the populations being

compared, the invariance property of item and ability parameters provides advantages over other statistical techniques when used for this purpose. However, these potential advantages are gained at a cost. One aspect of this cost is the complexity of working with these models. The many decisions that must be made when using item response models for this purpose are typically not straightforward. Decisions concerning model choice, model-data fit, test scaling, and detecting differences in item characteristic curves (ICCs) may not be straightforward and the outcome of these decisions can greatly affect the results of a translation equivalence study.

A particularly important and potentially troublesome aspect of using item response models to establish translation equivalence is the scaling of item and ability parameters from the examinee samples in the populations being compared. Item parameter and ability estimates obtained from different groups must be placed on a common scale before the ICCs from these groups can be compared. When establishing translation equivalence, common or "equivalent" items are necessary to accomplish this required scaling. One difficulty is that it is not known which of the test items should be used as common items. Furthermore, it is not clear from the test scaling literature how many common items are required to provide adequate scaling results. The problem of finding an anchor test of adequate length to provide adequate scaling of item parameter and ability estimates is an important issue in using item response models to establish translation equivalence. Without adequate scaling, the results of a translation equivalence study based on the use of item response theory (IRT) are suspect.

1.3 Purpose of the Dissertation

The previous section outlined several potential problem areas when using item response models to establish translation equivalence. Of particular concern is the scaling of item and ability parameters that are obtained separately from the examinee samples in the populations being compared. Therefore, the purpose of this investigation was to answer four questions:

1. How do differences in calibration sample size affect the anchor test length required to provide reasonably accurate IRT scaling results?
2. How do differences in the mean ability of examinee groups affect the anchor test length required to provide reasonably accurate IRT scaling results?
3. How does the interaction of these two factors affect the anchor test length required to provide reasonably accurate IRT scaling results?

And, finally

4. What anchor test length will provide reasonably accurate IRT scaling results?

1.4 Organization of the Thesis

The remainder of the thesis is organized into five chapters. Chapter 2 contains a review of the literature on test translations, the use of item response models in establishing translation equivalence and, test scaling through item response theory. Chapter 3 contains a discussion of the methodology used in this study. Chapter 4 contains the results of this study. Lastly, the conclusions from this study are presented in Chapter 5.

CHAPTER 2

REVIEW OF THE LITERATURE

2.1 Introduction

Adapting measurement instruments for use in a population other than those the instruments were designed for is a common practice (Samuda, 1975; Fouad & Hansen, 1987). Samuda (1975) reported that over 31 intelligence tests, aptitude tests and occupational/interest inventories have been adapted for this purpose. The more commonly used tests and inventories are often adapted for use in many populations. For example, the Self-Directed Search Interest Inventory has been translated into ten languages (Hansen, 1987).

Different procedures for adapting tests and inventories for use in other populations can be used. The adaptation procedure used depends on the degree of differences between the characteristics of the population for which the instrument was originally developed and on the population for which the adapted version is intended (these two populations are respectively termed the source and target populations). For example, if the source and target populations use the same language but exhibit cultural differences, an instrument can be adapted for use in the target population by relatively straightforward modification of items in the source version so they become more suitable for the target population (e.g., replacing the English number system with the metric number system in a mathematics item or substituting a term that is more relevant in the target population). However, if the source and target populations use different languages, the adaptation procedure used is language

translation of the source items. Since many source and target populations of interest use different languages, this chapter is focused on adapting tests through the use of language translations.

The purpose of this review is to provide an overview of language translation of tests and inventories, and the methods used to establish translation equivalence. The discussion that follows focuses on tests with the understanding that much of the discussion is generalized to occupational and interest inventories as well. The following topics are discussed in sections 2.2.1 to 2.2.5: (1) The Purpose of Test Translations, (2) Past and Present Trends of Test Translation Use, (3) Problems Associated with Test Translations, (4) Methods of Establishing Translation Equivalence, and (5) Examples of Translation Equivalence Studies. These sections are based on a review of the relevant literature.

2.2 Test Translations

2.2.1 The Purposes of Test Translations

Developing a test for use in a specific population can be accomplished by either (1) developing the test within the cultural boundaries of the population of interest, or (2) translating an existing test so that it is appropriate for the population of interest. If the purpose of developing a population-specific test is to reduce cultural bias in the test scores, either one of the development methods may be used; however, certain purposes require the use of the second method - test translation.

The first purpose that requires the use of test translation is the economical development of tests that are valid for use in specific populations or sub-populations. Some nations do not have sufficient

numbers of qualified personnel available for test development and validation. In such cases, translating existing tests is the only viable alternative for test development. For example, in a review of educational testing in Chile, Grassau (1969) stated: "A more important factor that prevents sound development in educational testing at this moment is the lack of personnel trained in testing." As a result of this lack of trained personnel, Chile has relied extensively on test translations to meet its country's testing needs (approximately 38% [n=21] of the intelligence tests used in Chile, as reported by Grassau [1969], were translated versions of existing tests).

A second purpose that requires the use of test translation is providing a basis for comparisons between populations (either distinct populations or within a population whose members' primary language or other cultural traits differ). For example, the development of a national test of mathematics achievement in India required developing equivalent tests in the thirteen regional languages spoken throughout the country (Kulkarni, 1969). Since the original mathematics test was written in English, thirteen test translations were required to produce a test with "equivalent" forms that could be used for national comparisons. A more recent example is the 1988 International Assessment of Educational Progress (IAEP). This assessment project required translating science and mathematics test items from English to French, Korean, and Spanish in order to make comparisons of achievement in these subjects across several populations (Lapointe, Mead, & Phillips, 1989).

While both purposes for test translations are valid, it is the second purpose - cross-population comparisons - that are of particular interest since test translations are the only alternative for allowing

such comparisons. Nations lacking sufficient numbers of qualified personnel for test development may have the option of acquiring more expertise, thus reducing the need for test translations; however, those involved in cross-population comparisons are more dependent on the use of translation techniques.

2.2.2 Past and Present Trends of Test Translation Use

The first test translated into another language was the Binet-Simon intelligence test. Henry Goddard translated the test from French to English in 1911 for use at the Vineland Training School for the mentally retarded in New Jersey (Stanley & Hopkins, 1972). Terman (1916) translated the original French version into English as part of the development of the Stanford-Binet intelligence test. By 1916, the Binet-Simon test had been translated into seven languages (Stanley & Hopkins, 1972).

Since these early test translations, numerous tests have been translated into the primary language of the examinees to be tested. Some examples include the Otis Group Intelligence Scale (1937), Wechsler Intelligence Scale for Children (1949), and the Wechsler Adult Intelligence Scale (1964). However, criticism of test translations has also paralleled the use of this technique. Pinter (1927) and Sanchez (1934) criticized the direct use of translated tests without first providing evidence of adequate validation (Swanson & Watson, 1982). Other critics have included Roca (1955), Quay (1971), Samuda (1975), Mercer (1979), and Perez (1980). Underlying much of the criticism were problems in (a) establishing equivalence in vocabulary, (b) determining the dominant language of target population examinees, and (c) cultural differences in responding to stimuli.

Despite these criticisms, tests (and questionnaires/inventories) are continually being translated for use in target populations (a list of translated tests, questionnaires, and inventories is provided in Appendix A). The reasons for this are clear. First, the development of population-specific tests for certain purposes (described in section 2.2.1) requires the use of test translations. Second, empirical studies support the use of test translations. Partial or total equivalence of translations have been reported by Brislin (1970); Katerburg, Hoy, and Smith (1977); Hulin, Drasgow, and Komocar (1982); Hansen and Fouad (1984); Hulin and Mayer (1986); Fouad and Hansen (1987); and Candell and Hulin (1987). For these two reasons, test translations have become an important aspect of test development work, particularly in the areas of intelligence and aptitude tests.

2.2.3 Problems Associated with Translating Tests

The use of tests in populations other than those the test was designed for has raised concerns since the beginnings of intelligence testing (Blanton, 1975; Samuda, 1983). In the case of test translations, it is assumed that enough differences between the populations of interest exist to warrant the development of a translated version of a test - it is identifying these differences and incorporating solutions to minimizing them that underlie many of the problems associated with translating tests.

Identifying and Minimizing Cultural Differences. An initial problem in the translation process is identifying the cultural differences between the source and target populations that may affect examinee test performance. Among these cultural traits are examinee motivation, values, experiences, and degree of test anxiety (Anastasi,

1954; Zirkel, 1972; Samuda, 1975; DeBlassie, 1988; van de Vijver & Poortinga, 1988). Cross-cultural researchers have provided numerous examples of how these cultural variables can influence the testing process. Van de Vijver & Poortinga (1988) point out difficulties experienced by Porteus in the administration of the Porteus Maze Test:

> . . . Porteus himself (1965) for instance, found it difficult to persuade Australian aboriginal subjects to solve the items by their own effort rather than in cooperation with the tester. As another example, it can be mentioned that the Maze Test, which is a paper-and-pencil test, has been applied among groups from which the members had never touched a pencil before. (Porteus, 1965, p. 3)

The same authors question the use of mazes as a suitable stimulus material for certain cultural groups:

> In the case of some cultural groups it is even debatable whether mazes are suitable as stimulus material. In a discussion on the use of the Maze Test among Bushmen, Reuning and Wortley (1973) argue that "the idea of a maze is not likely to occur to a Kalahari-dweller (like the Bushman) and must be utterly foreign to him" (p. 61). Their argument is based on the consideration that in a savannah, the natural ecology of the Bushmen, a person can invariably go along a more or less straight line from one point to another. (p. 3)

A third example is provided by Kline (1983), who discussed culturally related difficulties with using projective tests in certain populations:

> TAT (Thematic Apperception Test) and similar tests portraying figures or animals are culture bound, probably more so than psychometric test items. Lee (1953) attempted to produce an African TAT, but this proved suitable for few groups. Animals have deep cultural significance (e.g., Corman, 1966). Pigs raise considerable problems in Muslim or Jewish groups, and others have totem or taboo meanings for many groups. (p. 346)

Each of these examples, even though they do not deal directly with test translations, points out that cultural differences between the source and target populations can affect examinee performance. It is

therefore important to identify these cultural differences as a first step towards minimizing these effects. A further complication is that cultural differences must be considered for all components of the testing process including test instructions, test items (content, response format, response mode, and symbol usage), administrator-examinee interactions and testing environment (Berry & Lopez, 1977; van de Vijver & Poortinga, 1988).

Once identified, steps must be taken to minimize the impact of cultural differences on the testing process. For example, in a translation of a mathematics test written in English, it may be inappropriate to use an italicized x to represent an unknown angle in a geometry item for certain target populations. The use of this symbol in that context may not be culturally relevant and therefore a substitute unknown angle designation that is familiar to the target population examinees should be used. Failure to adequately control for cultural differences during the translation process can undermine the valid interpretation of the resulting test scores.

Identifying the Appropriate Language for Testing Target Population Examinees. A second area of problems associated with test translations is identifying the appropriate language to be used when testing examinees in the target population. Problems may arise because of varied dialects within the target language (Berry & Lopez, 1977; Olmedo, 1981). Olmedo (1981) noted: ". . . it is not uncommon to find that many tests written in formal Spanish are used inappropriately with populations that speak substantially different Spanish dialects." Unless examinees are being tested on their abilities with a formal language, at a minimum, even if translations to accommodate varied

dialects are not being done, it is important to identify the dialects spoken in the target language (and what members of the target population speak them) in order to make valid test score interpretations.

An even more complex problem associated with language and test translations is determining the most appropriate language for testing bilingual target examinees. DeAvila and Havassy (1974) pointed out that, because a person speaks a language, it can not be assumed that s/he can read and therefore be non-verbally tested in that language (neither can it be assumed that a person thinks in that language). Moreover, a person may only be a functionally receptive bilingual. For example, "children from homes where parents prefer to speak Spanish may themselves be only functionally receptive bilinguals. They may understand Spanish but express themselves in English. The situation with the parents may be the reverse" (Olmedo, 1981). These situations point out the importance of understanding the extent of bilingualism and its implications for testing in bilingual target examinees. Failure to determine the most appropriate language for testing the target population can seriously undermine the validity of translating a test from the start.

Finding Equivalent Words or Phrases. A third problem associated with language and test translations is finding, if they exist, words or phrases that are equivalent in the source and target languages. For example, in a Spanish translation of the Strong-Campbell Interest Inventory (II), Hansen and Fouad (1984) had difficulties finding an equivalent Spanish translation for the English word "argument" (the authors report similar difficulties with seven additional items). A second example was provided by Hulin, Drasgow, and Komocar (1982). They

had difficulty finding an equivalent Spanish translation for the English word "challenging" in the Job Descriptive Index.

Regional differences in word meaning within a single language can further complicate finding equivalent source to target language translations. DeAvila and Havassy (1974) used the Spanish word "toston" as an example; while "toston" is an appropriate translation for a "quarter" to speakers of Chicano Spanish, it means a "portion of a banana squashed and fried" to speakers of Puerto Rican Spanish. Consequently, even if equivalent words or phrases can be found, the assumption of translation equivalence must be checked for all sub-groups of the target population.

In an attempt to alleviate the problem of non-equivalent words or phrases in the source and target languages, a process known as decentering is sometimes used. Decentering refers to the modifying of words or phrases in either initially the source version of a test or later, in both language versions of a test in order to achieve item equivalence (Brislin, 1971). For example, the Spanish word "paloma" is equivalent to either "dove" or "pigeon" in English (Swanson & Watson, 1982) and therefore a test item in English that requires making a distinction between a dove and a pigeon would be difficult to translate into Spanish. The original item in English could be decentered by using a pair of terms that have similar meanings within the context of the item, and have equivalent terms in Spanish, thus allowing for a translation of the item.

Hulin and Mayer (1986) pointed out, however, that decentering may introduce psychometric nonequivalence between the original and translated item:

> Decentering produces translated material with smooth and natural terms in both versions. The price paid for such linguistic achievement may be that neither version is centered in either culture or language. Decentering should produce symmetrical translations with equal degrees of familiarity, colloquialism, and idiosyncrasy in both languages but fidelity to neither. The optimally decentered version, chosen through a mixture of back translations and discussions among translators (Brislin, 1980), may introduce serious questions about psychometric equivalence between the two versions. For instance, an English version of a questionnaire that contained the phrase "Once in a blue moon" (to describe the frequency of promotions) might result in a decentered Spanish phrase, "Every time a bishop dies." Linguistically and ethnographically, the two versions are equivalent. The price of linguistic smoothness, however, may be paid in the coin of psychometric nonequivalence.

Unfortunately, it is difficult to get a sense of the extent and appropriateness of decentering used in specific test translations from the literature; descriptions of test translations often report only whether decentering was used or not (an exception is Roca, 1955). Useful information for evaluating the decentering process might include the percentage of items decentered and illustrative examples of how the decentering was accomplished.

Finding Competent Translators. Lastly, there are also practical problems associated with test translations. Translators familiar with the source and target language and competent in the material covered by the source test can be difficult to find. Fink (1963) was unable to find translators competent in English and Laotian; consequently, a double translation from English to Thai and then from Thai to Lao was required (Brislin, 1970). The problem of finding competent translators becomes compounded when the test covers a specialized content domain (for example, medicine).

In summary, four problems associated with translating tests have been discussed. These include: (1) identifying and minimizing cultural

differences, (2) identifying the appropriate language for testing the target population examinees, (3) finding equivalent words or phrases, and (4) finding competent translators. The extent to which each of these is a problem in translating a test will, of course, vary depending on the characteristics of the test and of the source and target populations. For example, it may be more difficult to identify and minimize cultural differences for a test with a high degree of verbal loading than a test that makes greater use of symbols. Moreover, if the characteristics of the source and target populations differ greatly, identifying and minimizing cultural differences will be more difficult than for source and target populations with similar or overlapping characteristics. Translating a test from one language to another and maintaining its validity with respect to a specific purpose can be an exceedingly complex process. Being aware of the many potential problems in translating tests may help to minimize the errors associated with the translation process.

2.2.4 Methods of Establishing Translation Equivalence

Equivalence of test items is defined as the direct comparability of test items and the scores derived from them in terms of psychometric meaning. Thus, test items are equivalent if they measure the same behaviors across the populations of interest and examinees with equal amounts of ability within the populations have equal probabilities (within the limits of measurement error) of answering the items correctly.

A review of the literature on test and inventory translations indicated that many different methods have been used to establish the equivalence between source and translated instruments. Some of the

methods are more commonly used than others; however, a comprehensive review of most or all of the available methods seemed useful. These methods include those that are used both before and after examinee responses have been collected. Each of the methods will be discussed mostly in terms of tests and test items with the understanding that these discussions generally apply to questionnaires and inventories as well.

The methods of establishing equivalence between original and translated test items can be viewed as an extension of the methods used for identifying item bias. In bias studies, the focus is on the items or scores derived from them for a single test. Establishing translation equivalence extends this focus to the items or scores derived from them on two tests - the original test and either the initial translation or the back translated version of the original test. The presence of more than one version of a test on which to compare scores gives rise to the various methods of establishing translation equivalence to be discussed.

There is also a similarity in the methods used to establish translation equivalence and to identify biased items. In each case, both (a) judgmental and (b) statistical methods may be used. Judgmental methods of establishing translation equivalence are based on a decision by an individual or a group on the degree of each item's translation equivalence. In contrast, statistical methods establish translation equivalence based on the analysis of examinee responses to some combination of the original, translated, or back translated test items. The use of judgmental and statistical methods is not necessarily independent. Judgmental methods are often used as preliminary checks of

translation equivalence before the tests are administered and statistical methods applied to the test scores.

The classification scheme adopted for identifying methods of establishing translation equivalence in this study is based on whether judgmental or statistical methods are used. In addition, it is also useful to identify whether a single or back translation is used. Therefore, four categories of methods can be identified:

1.A Judgmental single-translation methods

1.B Judgmental back-translation methods

2.A Statistical single-translation methods

2.B Statistical back-translation methods

Figure 1 provides an overview of the current methods within each of these categories.[1] These methods (7 total) are discussed next. The specific statistical techniques, such as factor analysis or analysis of variance that can be used with the statistical methods (2.A.1 to 2.B.1) will also be discussed.

Judgmental Methods. As stated previously, judgmental methods of establishing translation equivalence are based on a decision by an individual or a group on the degree of each item's translation equivalence. Thus, judgmental methods provide a subjective viewpoint on the question of equivalence (even though statistical procedures can be applied to help in evaluating the validity and reliability of judgments or ratings, the basic source of information is individual or group judgments). It is worth repeating that each of these judgmental (and

[1]References to method 2.B.1 were not found, indicating that either it has not been used or is not a popular method of establishing translation equivalence. It is presented here because the design is interesting and seems potentially useful.

the statistical) methods need not be considered in isolation. Any combination of methods can be, and often are, used. For example, it would be unusual to find statistical procedures being used to establish equivalence without some type of judgmental method being used first. Furthermore, multiple judgmental or statistical methods are often used.

In addition, judgmental methods are appropriate for establishing translation equivalence of those aspects of a test for which scores cannot be obtained and for which statistical methods are not applicable (e.g., test instructions and orientation materials).

Post-translation probes. In this method (1.A.1), one or more samples of target examinees answer the translated version of an item and are then asked about the meaning of their answers (Brislin, 1970). Evidence of translation equivalence is obtained if the responses given by a high percentage of the examinees questioned reflect a reasonable interpretation of an item in terms of cultural and linguistic understanding. The main judgmental aspect of this method is deciding what responses by target examinees about the meaning of their answer to an item are considered reasonable.

The use of this method can provide valuable insights into why an item did not successfully translate since examinees can be directly asked about their interpretation of an item. This advantage can, however, be offset by the interaction between the prober and the examinee being questioned. Cultural, linguistic, and possibly personality differences between the prober and examinee can interfere with the results obtained from the post-translation probe.

1. Judgmental Methods

1.A Judgmental single-translation methods

	Source	Target
1.A.1.	---------->	Post-translation probes
1.A.2	---------->	Bilingual judges check errors
1.A.3	---------->	Performance criteria - perform a task using translated instructions

1.B Judgmental back-translation method

1.B.1 ---------->
<----------

Source language monolinguals check for errors

2. Statistical Methods

2.1 Statistical single-translation methods

	Source	Target
2.A.1	---------->	Bilinguals take source and target versions
2.A.2	---------->	
	Source language monolinguals take source version	Target language monolinguals take target version

2.B. Statistical back-translation method

2.B.1 ---------->
<----------

Source language monolinguals take source and back-translated versions

Figure 1. Methods for Establishing Equivalence of Translated Test Items

A second problem with this method is that it is relatively labor intensive compared to many other judgmental methods. In addition to enlisting and using probers, examinees are needed to answer test items and respond to probes. Additionally, the probing process is likely to be a time-consuming one.

A third problem with this method is that one has to be sure of the meaning of the answers from source language monolinguals in order to judge the equivalence of the meaning of answers from target language monolinguals. In other words, the validity of the test in the source population must be fully checked before comparing results from source and target examinees. Although any source language test should be validated before it is translated, it becomes particularly important when using this method since no comparison to the source version is being made. For tests that have not undergone stringent validity checks in the source population (for example, tests that have been developed for small scale research studies), it may be useful to probe a sample of source language monolinguals as well. This sample of monolinguals should be matched as closely as possible to target examinees on the ability or abilities of interest. With this additional check, the problem of comparing irrelevant scores can possibly be avoided.

Bilingual judges check for errors. Method 1.A.2 makes use of bilingual judges who compare the source and translated versions of each test item and decide whether any differences between translations could result in non-equivalence of meaning in the two populations of interest (Brislin, 1970). These comparisons can be made on the basis of having judges simply look the items over, check the characteristics of the items against a checklist of item characteristics that may introduce

non-equivalence, or by having them attempt to answer both versions of the items before comparing them for errors.

One problem in applying this method is that it is often difficult to find bilingual judges who are equally familiar with the source and target languages and/or cultures (this issue was raised previously in the context of determining the most appropriate language for testing bilingual target examinees).

> . . . the judgment equally bilingual and bicultural is extremely difficult, perhaps even impossible, to make. More than likely, the individual members of the group, and even the group as a whole, will on average be more proficient in one of the two languages than in the other. This will be especially true, of course, if the group is small. (Angoff & Cook, 1988)

Therefore, judgments about differences between the source and translated versions are subject to variations from this source of error.

A second problem with this method is that bilingual judges may inadvertently use "insightful guesses" to infer equivalence of meaning. This problem is usually raised in the context of using back-translation techniques. Hulin (1987) noted:

> Apparently equivalent terms, such as *amigo*, *friend* and *tovarish*, are not always equivalent, but translators sharing a small number of rules-of-thumb may consistently translate such terms as if they were equivalent. Equivalent source language versions may be generated from poorly translated and constructed target language versions by insightful guesses and assumptions by the translators about what the term must have meant in the original language. Translations that retain grammatical forms of the original language are easy to back-translate but may not be meaningful to target language monolinguals (Brislin, 1970).

Judges are also translators of a sort and are subject to the same errors, in this case using "insightful guesses" to infer equivalence of meaning, as those who performed the initial translation.

A third problem with this method is that bilingual judges may not think about an item in the same way as their respective source and target language monolinguals.

> Bilingual individuals have cognitive and semantic structures that may differ from the structures of either group of monolingual individuals with whom they share a language (Ervin-Tripp, 1964; Mannamara, 1970; Peal & Lambert, 1962; Segalowitz, 1980). (Hulin, 1987)

Consequently, the use of bilingual judges to establish translation equivalence may lead to results that are not generalizable to source and target language monolinguals. This problem raises serious questions about the overall usefulness of this method for establishing translation equivalence.

Performance criteria. This method (1.A.3) of establishing translation equivalence is based on the criterion that "if people could perform bodily movements after having heard either a source or target language instructions, and if the results of the bodily movement criterion were similar across all people, then the source and its translation must be equivalent" (Brislin, 1970). The obvious limitation of this method is that it can only be used with testing materials that can be evaluated through bodily movements such as some test instructions or performance test items. This limitation and the fact that this method, like method 1.A.1 - Post-translation probes - is also relatively (1) labor intensive and (2) sensitive to prober-examinee interactions, reduces the general usefulness of this method for establishing translation equivalence.

Source language monolinguals check for errors. Back translation refers to the translation of the target version test back into the source version by bilinguals not involved in the original translation in

order to check for translation equivalence (Brislin, 1970). Translation equivalence using this method (1.B.1) is established by having source language monolinguals check for errors between the source and back-translated versions of a test (Brislin, 1970; Hulin & Mayer, 1986; Hansen, 1987).

The main problem associated with the use of this method is the reliance on the assumption that errors made during the original translation will not be made again (in reverse) during back-translation. However, as discussed in relation to method 1.A.2, a translator may use "insightful guesses" or "rules-of-thumb" to translate an item, thus making it appear equivalent to the source item even though it may not be (if this were not the case, checks on the equivalence of the original translation would not be as necessary). Likewise, the use of these "insightful guesses" and "rules-of-thumb" during the back-translation process can mask those errors made during the original translation. Brislin (1970) reported finding errors due to translation after three successive translation/back-translation sequences, indicating that the assumption that the same errors that occurred in the original translation will not occur, in reverse, during back translation is questionable. The use of additional (independent) translators may make it more likely that differences in the original translation will be detected, but the high potential for the violation of the previously mentioned assumption reduces the usefulness of this technique and <u>any</u> of the methods discussed that are based on its use.

This is not to say that back-translating is not a useful technique; rather, that it should be considered a general check on translation quality that will most likely detect obvious errors in the

original translation. For example, in an effort to establish translation equivalence of a Spanish translation of the Job Descriptive Index, Hulin, Drasgow, and Komocar (1982) used the back-translation technique as an initial check of translation quality before applying another method of establishing translation equivalence:

> Translation of psychological scales into new languages involves a series of steps. First, translation into the target language and back translation into the original language by multiple independent translators is required. This is simply a check and verification on the general quality of the translation and should be done for any translation. Lack of convergence back into the original language is apparent. Remedial action can be achieved at this point by refining problem items. Back translation is necessary but not sufficient for generating equivalent scales. (Hulin, Drasgow, & Komocar, 1982)

Other examples of the use of the back translation technique as an initial check of translation quality are provided by Hansen (1987) and Katerburg, Smith, and Hoy (1977).

Statistical Methods. The various statistical methods to be discussed result from variations in the type of examinee responding (source language monolinguals, target language monolinguals, or source-target bilinguals) and the version of the test (original, translated, or back-translated) responded to. Altogether, four statistical methods will be discussed. In order to facilitate the discussion of the statistical methods of establishing translation equivalence, the potential statistical techniques used with the four statistical methods (2.A.1 to 2.B.1) will be introduced first.

The statistical techniques used with the various methods of establishing translation equivalence can be categorized along two

dimensions.[2] The first dimension is whether it is assumed that the test constitutes a common scale on which scores can be compared. The second dimension is whether the statistical technique conditions on the ability of the examinees to be compared. Using these two dimensions, Table 1 was formed.[3] Examples of statistical techniques for each of the cells are given.

Table 1

Classification of the Statistical Techniques Used to Establish Translation Equivalence

Scale	Technique
Common Scale Not Assumed	Factor analysis, comparison of correlation matrices
Common Scale Assumed (Unconditional)	Analysis of variance, analysis of p-values or transformed item difficulties
Common Scale Assumed (Conditional)	Item response models, chi-square approaches

The following is a more complete list of these statistical techniques based on the classification scheme given in Table 1. The citations provided are those authors who either mentioned or used that statistical technique. If the author(s) used the technique, the citation is underlined.

[2]This classification scheme is adapted from van de Vijver and Poortinga (1991).

[3]Table 1 is adapted from van de Vijver and Poortinga.

Common Scale Not Assumed:

A. Factor Analysis - Exploratory/Confirmatory (Irvine & Carroll, 1980; Kline, 1983; Poortinga, 1983; Mayberry, 1984; Hulin & Mayer, 1986; Van de Vijver & Poortinga, 1991).

B. Comparison of correlation matrices (Poortinga, 1983; Van de Vijver & Poortinga, 1991).

Common Scale Assumed (Unconditional)

A. Correlation between scores (Hansen & Fouad, 1984; Hulin & Mayer, 1986).

B. Item-total correlations (Poortinga, 1983; Hulin, 1987).

C. Response frequency of distractors (Poortinga, 1983).

D. Analysis of variance (Katerburg, Smith, & Hoy, 1977; Irvine & Carroll, 1980; Kline, 1983; McCauley & Colberg, 1983; Poortinga, 1983).

E. Generalizability theory (Katerburg, Smith, & Hoy, 1977; Kline, 1983; Hulin, 1987; Van de Vijver & Poortinga, 1991).

F. Correlation between transformed p-values (McCauley & Colberg, 1983; Poortinga, 1983; Hulin, 1987).

G. Plots of transformed p-values (Hulin, 1987; Van de Vijver & Poortinga, 1991).

H. Comparison of mean scores (van der Flier, 1982; Hansen & Fouad, 1984; Hulin & Mayer, 1986).

I. Comparison of standard deviations (Hulin & Mayer, 1986).

J. Correlation between individual scores (Hansen & Fouad, 1984).

Common Scale Assumed (Conditional)

A. Item response models (Irvine & Carroll, 1980; Hulin, Drasgow, & Komocar, 1982; van der Flier, 1982; Kline, 1983; Poortinga, 1983; Hulin & Mayer, 1986; Candell & Hulin, 1987; Hulin, 1987; van de Vijver & Poortinga, 1991; Simon, 1989).

B. Chi-square analysis (Kline, 1983; Poortinga, 1983; Van de Vijver & Poortinga, 1991; Simon, 1989).

C. Partial correlation (Simon, 1989).

D. Mantel-Haenszel analysis (Simon, 1989).

E. Iterative logit procedure (Simon, 1989).

Two comments concerning these statistical techniques are in order. First, as van de Vijver and Poortinga (1991) point out, the distinction between the conditional and unconditional statistical techniques is not absolute but rather is dependent on the empirical use of a particular technique:

> . . . the classification of particular techniques as unconditional methods is mainly determined by their empirical use. The [unconditional] methods mentioned can also be applied as conditional methods, namely by including level of ability as an additional factor in the analysis. Suppose a researcher wants to compare p-values obtained in various cultural groups. An unconditional analysis entails a direct comparison of the item statistics, while in a conditional analysis the samples of subjects will be divided according to the level of their raw score and analyzed per level. Conversely, the conditional methods which will be discussed, can also be used in an unconditional way by eliminating ability as a separate factor during the analysis. (van de Vijver & Poortinga, 1991)

Second, more than one statistical technique is often used with a statistical method of establishing translation equivalence. For example, to establish the degree of translation equivalence for the English to Spanish translation of the Strong-Campbell Interest Inventory, Hansen and Fouad (1984) used the following statistical techniques in conjunction with method 2.A.1 (bilinguals take source and target versions): (1) correlation between group scores on the two forms and (2) comparison of the mean scores on the two forms. Another example was Candell and Hulin's (1986) use of factor analysis to assess the dimensionality of the scores derived from the English and Spanish versions of the Job Descriptive Index before applying an item response model (with method 2.A.2 - source language monolinguals take source

version and target language monolinguals take target version) to establish translation equivalence.

Bilinguals take source and target versions. In this method (2.A.1), bilingual examinees take both the source and target versions of a test (with an adequate time interval in between administrations) and the scores on the two tests are then compared (Brislin, 1971; Katerburg, Smith, & Hoy, 1977; Hulin, Drasgow, & Komocar, 1982; Hansen & Fouad, 1984; Candell & Hulin, 1986). The source version of the test can either be the original version or a version that has been revised after being checked for translation equivalence with another method. The appeal of this method is that by having the same examinees take both versions of a test, differences in examinee ability that can confound translation equivalence will be controlled for. However, the problem of unequal examinee bilingualism and/or biculturism (discussed with method 1.A.2 - using bilingual judges to check for errors) also applies to the examinees used with this method. The possibility of unequal examinee bilingualism and/or biculturism can violate the assumption of equal examinee ability. Therefore, the assumption that the use of bilinguals controls for differences in ability that would most likely occur if separate source and target language monolinguals were used instead is questionable.

One way to strengthen this method is to use examinees who are identified as being equally bilingual by a test of language dominance. For example, English-Spanish bilingual examinees could be tested using the Flexibility Language Dominance Test or the Bilingual Syntax Measure and those examinees whose scores indicate that they are equally (or

close to equally) bilingual would then take the source and translated version of a test. Several drawbacks with this additional step are evident. These include (1) obtaining or developing a test of language dominance for the source and target languages of interest, (2) the additional required testing time, and (3) the lack of counterpart tests that address biculturism or culture dominance. This additional step may, however, be a practical addition to this method when a test of language dominance appropriate to the source and target languages is readily available (for example, tests of language dominance for English-Spanish are readily available in the United States).

Another way to strengthen this method is to use statistical techniques that condition on examinee ability. In the few examples provided in the translation literature where this method of establishing translation equivalence was used, unconditional statistical techniques such as correlations between scores or the use of generalizability theory have been used to compare examinee scores from the source and target versions of the test. These unconditional statistical techniques were used because it was assumed that the use of bilinguals controls for differences in examinee ability. However, as previously mentioned, this assumption is questionable and therefore the use of conditional statistical techniques, such as the use of item response theory, can be used to strengthen this method of establishing translation equivalence.

Another comment concerning the use of bilinguals in establishing translation equivalence deserves mention. Historically, bilingualism was thought to be a language handicap that interfered with intellectual development and academic achievement (see reviews by Darcy, 1953, 1963). In contrast, recent research in this area (see review by Diaz, 1983)

indicates that compared to monolinguals, bilinguals who are equally proficient in the use of two languages "show definite advantages on measures of metalinguistic abilities, concept formation, field independence, and divergent thinking skills" (Diaz, 1983). Thus, in using bilinguals to establish translation equivalence, the resulting scores may be in general higher than if source and target language monolinguals were used. In the extreme case, floor effects may be noted when the final version of the source and target tests is administered to monolinguals in their respective languages. This problem can arise due to errors in sampling as well, but the use of bilinguals can possibly add a further dimension to this source of error.

The most serious problem with this method, however, is that the scores obtained from bilingual examinees may not be generalizable to their respective source language monolinguals (this problem was raised with method 1.A.2 - bilingual judges check for errors). This problem has been tested empirically by Drasgow and Hulin (1986). They compared previous results of establishing translation equivalence of a Spanish translation of the Job Descriptive Index where bilingual subjects were used (Hulin, Drasgow, & Komocar, 1982) to results using monolingual subjects. In both cases, item response models were used to establish translation equivalence. When bilingual subjects were used (method 2.A.1), approximately 4% (3 out of 72) of the items were determined to have been poorly translated as compared to 30% when monolingual samples (method 2.A.2) were used. Hulin and Mayer (1976) conducted a similar study and obtained similar results. These discrepancies in the number of items identified as poorly translated indicates that the results of

establishing translation equivalence based on bilingual responses are likely not generalizable to monolingual populations.

This problem of generalizing results from bilinguals to monolingual populations has been the major impetus for the increased interest in method 2.A.2 (source language monolinguals take source version and target language monolinguals take target version). With the use of method 2.A.2, samples from the two sub-populations we are interested in generalizing to (source and target monolinguals) are used and questions of generalizability are relegated to the choice of sample used.

Source language monolinguals take source version and target language monolinguals take target version. In this method (2.A.2), source and target language monolinguals are used, with each taking the version that is in their respective languages (Candell & Hulin, 1986; Hulin & Mayer, 1986; Hulin, 1987). The source version of the test can either be the original version or a version that has been revised after being checked for translation equivalence with another method. The two sets of scores are then compared to establish the extent of translation equivalence between the two versions.

The main advantage of this method is that source and target language monolinguals are used and therefore the results of establishing translation equivalence based on this method are more generalizable to these two sub-populations than the statistical methods that use only source language monolinguals (2.B.1) or, to a lesser extent, bilinguals (2.A.1) as examinees. This is due to the concern that bilinguals may not respond to items in the same way that monolinguals in either language do (this problem was raised with methods 1.A.2 - bilingual

judges check for errors - and 2.A.1 - bilinguals take source and target versions) and, that using only source language monolinguals (method 2.B.1) necessarily precludes obtaining results from target language monolinguals. The use of source and target language monolinguals reduces the question of generalizability of the results obtained with this method to the choice of samples and the statistical techniques used.

The problem with this method is that two samples of examinees are used and therefore the resulting scores may be confounded with differences in ability between the two samples. However, alternative steps can be taken to minimize this problem.

First, in choosing samples of source and target language monolinguals, every effort should be given to matching examinees in the two groups on the ability or abilities or interest. An external criterion such as IQ or other test scores that are correlated with the tasks of interest may be available for this purpose. Alternately, if an external criterion is not available, examinee samples should be chosen using the most available information about the ability level of each sample. Information such as years and type of schooling, age, gender and demographic data may be used for this purpose.

Second, conditional statistical techniques that take into account the ability of examinees when comparing test scores can also be used to control for ability differences in the source and target examinee samples. Examples of conditional statistical techniques that can be used for this purpose include those based on the Chi-square statistic (Scheunemann, 1979; Shepard, Camilli, & Averill, 1981) and item response models (Lord, 1980).

The use of item response models are, in particular, receiving much recent attention as a statistical technique used with this method (Irvine & Carroll, 1980; Hulin, Drasgow, & Komocar, 1982; van der Flier, 1982; Poortinga, 1983; Hulin & Mayer, 1986; Candell & Hulin, 1987, 1987; Hulin, 1987; van de Vijver & Poortinga, 1991; Simon, 1989). The advantages of using item response models for this purpose will be discussed in section 2.3.1.

Lastly, factor analysis, or other statistical techniques in which no common scale for scores from the populations is assumed, is often used in conjunction with this method to establish translation equivalence (Irvine & Carroll, 1980; Kline, 1983; Poortinga, 1983; Mayberry, 1984; Hulin & Mayer, 1986; van de Vijver & Poortinga, 1991). In the case of factor analysis, scores from source and target language monolinguals are separately analyzed to determine the similarity of factor structures across the populations. A dominant first factor would provide evidence that the same underlying construct is being measured in the two populations by the source and target versions of a test. However, aside from the methodological difficulties associated with its use (for example, the choice of correlation coefficient used), the results of a factor analysis are limited in generalizability to similar samples of source or target language monolinguals. This is the case since factor analysis is based on classical item statistics and therefore the results are not sample invariant.

Factor analysis may be used in conjunction with this method when an item response model is to be used. In this case, its purpose is to check on the unidimensionality of the item responses within the two samples so that a unidimensional item response model can be used.

Source language monolinguals take original and back-translated versions. In this method (2.B.1), source and back-translated versions are both taken by source language monolinguals and, as with all of the statistical methods, the scores are then compared using one or more statistical techniques to establish the extent of translation equivalence. The advantage of using this method is that by using one sample of examinees, the resulting scores are not confounded with differences in examinee ability.

One problem with this method is that one set of scores is based on a back-translated version which, as discussed with method 1.B.1, can mask errors made during the original source to target version translation. An additional problem with the use of this method is that target language monolinguals are not used and yet, in part, we are attempting to generalize the meaning of the resulting test scores to a population of target language monolinguals. Making such generalizations without obtaining test scores from at least a sample of the population of interest appears to be a valid concern with the use of this method (and with method 2.A.1 which makes use of bilinguals although to a lesser extent).

No references to the use of this method could be found indicating that this method is either unpopular or has not been used. It is included here because this method appears to be a logical and practical extension of method 1.B.1 (source language monolinguals check for errors) which has been and is presently a popular method of establishing translation equivalence.

The discussion of the methods of establishing translation equivalence has so far focused on introducing the individual methods and presenting the advantages and problems of using each of the methods. What is evident from these discussions is that certain general problems with using the individual methods of establishing translation equivalence cross several of the methods. In an attempt to provide a basis for choosing one or more methods over others, six general problems will be briefly reviewed next.

1. Generalizability to the task of interest

We are ultimately interested in how examinees in the two populations of interest respond to the test items in their respective languages. A problem with method 1.B.1 (source language monolinguals check for errors) is that examinees are not required to answer test items (only to check for errors). Since comparing test items for errors in translation may involve different cognitive processes than responding to them, it may be incorrect to generalize from the task of checking for errors in test items to the task of responding to test items. This problem may also apply to method 1.A.2 (bilingual judges check for errors) when judges are asked only to compare source and target items instead of basing their comparison on their own responses to the items.

2. Generalizability to the populations of interest

A problem with methods 1.B.1 - source language monolinguals check for errors - and 2.B.1 - source language monolinguals take source and back-translated versions is that target language monolinguals are not used and yet it is this population that we

are, in part, generalizing the meaning of the resulting test scores to.

The same problem exists for those methods that make use of bilinguals (1.A.2 - bilingual judges check for errors - and 2.A.1 - bilinguals take source and target versions). In these methods, the assumption is made that bilinguals will respond to an item in the same way as monolinguals in either language. This is a questionable assumption to make and therefore it may confound the results obtained using these methods. However, the use of bilinguals will most likely be less of a problem in generalizing to the populations of interest than the use of only source language monolinguals.

3. Differences in judges' or examinees' ability

Method 2.A.2 (source language monolinguals take source version and target language monolinguals take target version) makes use of source and target language monolinguals and therefore the results obtained from this method may be confounded with ability differences between the two groups. This problem also applies to the methods that make use of bilingual judges or examinees (1.A.2 - bilingual judges check for errors - and 2.A.1 - bilinguals take source and target versions), although probably to a lesser extent than with the use of source and target language monolinguals. However, differences in group or bilinguals' abilities when using methods 2.A.2 or 2.A.1 can be controlled for by the use of conditional statistical techniques. The problem still remains with method 1.A.2, which uses bilingual judges,

since differences in judges' abilities between the source and target languages cannot be controlled for statistically.

4. Use of back-translations

The use of back-translations may cause problems in establishing translation equivalence because errors made in the original source to target translation may be made (in reverse) during the back translation (this may be due to insightful guesses made by the back-translator[s]). Thus, errors made in the original translation may be masked by using those methods that make use of back-translations (1.B.1 - source language monolinguals check for errors - and 2.B.1 - source language monolinguals take source and back-translation versions). Back-translating may be useful for picking up obvious errors in the original translation; however, it may not be as useful for picking up more subtle translation errors.

5. Sensitivity to examiner/prober-examinee interaction

All of the statistical methods require administering a test to examinees and, therefore, examiner-examinee interactions may effect the resulting scores. However, the judgmental methods that make use of post-translation probes (1.A.1) or performance criteria (1.A.3) are especially sensitive to examiner/prober-examinee interactions since these methods, in all likelihood, involve a high degree of contact between those administering the test or probes and examinees.

6. Labor intensive

Methods 1.A.1 (post-translation probes) and 1.A.3 (performance criteria) can be relatively labor intensive compared

to, for example, having bilingual judges check for errors (1.A.2). This will be particularly true if a large sample of target language examinees is used.

These six problems, and the methods of establishing translation equivalence to which they apply, are shown in Table 2. Besides providing an overview of the general problems associated with each method, this Table can be used to help minimize the errors associated with establishing translation equivalence when more than one method is used. For example, within the judgmental methods, it can be seen from Table 2 that methods 1.A.2 and 1.B.1 have two general problems in common and therefore these two methods should possibly not be used together to establish translation equivalence. A better combination to use may be methods 1.A.1 and 1.A.2 or 1.A.1 and 1.B.1 since these combinations do not share the same general problems. Across the judgmental and statistical methods, methods 1.A.3 and 2.A.2 may be a good combination to use for the same reason. Using more than one method will result in a more stringent check of translation equivalence when the methods used minimize the general problems they have in common.

However, the choice of method or methods should not be made simply on the number of problems avoided by their use. For one, some problems may be considered more serious than others. For example, budget or time limitations may rule out the use of those methods that are labor intensive (1.A.1 and 1.B.1). Even across methods, the seriousness of a problem may vary. An example is problem 2 (generalizability to the populations of interest), which is most likely a more serious problem when only source language monolinguals (1.B.1 and 2.B.1) rather than bilinguals (1.A.2 and 2.A.1) are used. External factors can also

Table 2

General Problems Associated with the Methods of Establishing Translation Equivalence

	Methods of Establishing Translation Equivalence						
	Judgmental Methods				Statistical Methods		
	1.A.1	1.A.2	1.A.3	1.B.1	2.A.1	2.A.2	2.B.1
Problem 1: Generalizability to the task of interest		X		X			
Problem 2: Generalizability to the populations of interest		X[1]		X	X[1]		X
Problem 3: Differences in judges' or examinees' ability		X			X[2]	X[2]	
Problem 4: Use of back-translations				X			X

Continued on the next page.

Table 2, continued:

	Methods of Establishing Translation Equivalence						
	Judgmental Methods				Statistical Methods		
	1.A.1	1.A.2	1.A.3	1.B.1	2.A.1	2.A.2	2.B.1
Problem 5:							
Sensitivity to examiner-examinee or prober-examinee interactions	X		X		X[3]	X[3]	X[3]
Problem 6:							
Labor intensive	X		X				

An X indicates the problem is associated with the method.

[1]Most likely less of a problem than using only source language monolinguals.

[2]Less of a problem if conditional statistical techniques are used with the method.

[3]Most likely less of a problem than when probes or performance criteria are used.

influence the seriousness of a problem. An example is problem 3 (differences in judges' or examinees' ability), where the seriousness of this problem for the statistical methods (2.A.1 and 2.A.2) varies depending on whether conditional statistical techniques are used with these methods or not. These examples point out that the choice of method or methods used depends on many factors. Table 2 can provide a frame of reference for considering the various available methods and potentially viable combinations, but the final choice of method or methods used should ultimately be based on further considerations as well.

An additional use for Table 2 might be to compare judgmental and statistical methods in identifying items that failed to translate well. This has been an important line of research in the study of item bias because identifying why judgmental methods failed to flag the same items as the statistical methods can lead to insights into the nature of item bias. This information can be used by item writers in reducing the number of biased items written and to help in developing better judgmental methods so potentially biased items can be detected before being administered to examinees. Likewise, comparing judgmental and statistical methods in identifying items that failed to translate well can provide comparable information and advantages in the context of translating test items.

Table 2 can be used when comparing judgmental and statistical methods for flagging poorly translated items by noting the number of problems shared by the judgmental and statistical methods being compared. If the two (or more) methods do not have some problem or problems in common, it would not be surprising to find inconsistent

results across the methods. An example would be comparing the judgmental method 1.B.1 with the statistical 2.B.1. Different problems have been identified across the two methods and therefore consistent results across the methods would appear unlikely from the outset. Similarly, the information in Table 2 could also be used when comparing across just judgmental or statistical methods. However, users are cautioned against interpreting Table 2 without considering other factors that may influence the seriousness of the problems mentioned.

In summary, seven methods (four judgmental and three statistical) of establishing translation equivalence have been introduced in this section along with a discussion of their respective advantages and problems. With the exception of method 2.B.1 (source language monolinguals take source and back-translated versions), these methods represent the methods of establishing translation equivalence that were found in a review of the relevant literature. Other methods are possible. For example, method 1.A.1 (post-translation probes) could be extended to include post-translation probes of source language monolinguals who take the source version of a test. Method 1.A.3 (performance criteria) could be extended in a similar way, resulting in an additional method of establishing translation equivalence. However, these additional methods are either variations or extensions of the basic methods presented here and, as such, their respective advantages and problems can be evaluated using the discussions presented in this section.

2.2.5 Examples of Translation Equivalence Studies

Two examples of studies to establish translation equivalence will be presented in this section in order to provide an overview of how the

methods of establishing translation equivalence (presented in section 2.2.4) have been used in practice. These two examples were chosen because together they illustrate the use of three of the more popular methods of establishing translation equivalence.

The first example was a study to establish the translation equivalence of the English to Spanish translations of the Job Descriptive Index (JDI) and the Index of Organizational Reactions (IOR) (Katerburg, Smith, & Hoy, 1977). The English versions of both instruments were initially translated into Spanish and then subsequently back-translated into English by translators who were not involved in the original translation. Method 1.B.1 (source language monolinguals check for errors) was then used to check for errors between the original and back-translated versions of the two instruments. Differences in meaning of words or phrases between the two versions highlighted some of the problems with the original translations. Translators then either (a) revised the Spanish version by attempting to find words or phrases that better matched the meaning of corresponding words or phrases in the original version or (b) decentered the English and Spanish versions so that words or phrases that are equivalent across both language versions could be used.

Method 2.A.1 (bilinguals take source and target versions) was used in the next phase of this study. Using a completely counterbalanced design (two language versions x two times), bilinguals were administered the two instruments in one of eight unique orders with a six-week interval between administrations. The resulting examinee responses were broken down by sub-scales and analyzed using generalizability theory resulting in generalizability coefficients and proportions of variance

due to time, language version, and person (and their interactions) for each of the sub-scales in the two instruments. Coefficients of stability for the English and Spanish versions of both instruments were also reported.

The second example of a study to establish translation equivalence is Angoff and Cook's (1988) study on the equating of the English and Spanish versions of the Scholastic Aptitude Test (SAT). Their study focused on (1) establishing the translation equivalence for a set of anchor items to be used in equating the two language versions of the SAT and (2) the equating procedure itself. Since we are mainly interested in the methods and procedures used to establish translation equivalence, the equating portion of this study will not be discussed here.

The first step in establishing translation equivalence was to translate the already existing English version of the SAT into Spanish and the already existing Spanish version into English. The two translated versions were then back-translated into their respective original languages by translators who were not involved in the initial translations of the two language versions of the test. Method 1.B.1 (source language monolinguals check for errors) was then used to check for errors between the source and back-translated versions for the two language versions of the test.[4] In each case, differences between the source and back-translated versions were noted and either (1) adjustments in the original translations were made if it was determined that the adjustments were adequate to provide potential translation

[4]Comparisons between English source and Spanish target (i.e., English translation of the original Spanish version) or between Spanish source and English target (i.e., Spanish translation of the original English version) were not mentioned by the authors.

equivalence or (2) the items were dropped as potential anchor items if it was determined that translation equivalence was unlikely to be obtained for these items.

The next phase of this study made use of method 2.A.1 (source language monolinguals take source version and target language monolinguals take target version). In this case, either the English or Spanish version can be considered the source or target version. After examinee responses from a sample of source and target language monolinguals were obtained, item characteristic curves (ICCs) were estimated separately for each of these groups (the three-parameter logistic model was used). The item parameters were then scaled to allow for comparisons of the ICCs between the two groups. The final set of ICCs for each group were obtained after using a criterion purification procedure developed by Lord (1980, chap. 14). This procedure reduces the problem of using ability and item parameter estimates that may be obtained from non-equivalent items to establish the equivalence of translated items (the steps used in this procedure will be discussed in section 2.3.4). The final set of ICCs for source and target language monolinguals was compared to establish the translation equivalence of potential anchor items that were to be used in equating the two language versions of the SAT.

Comparisons of ICCs were based on a combination of indices. First, a chi-squared item bias statistic was calculated for each item. This statistic tests the null hypothesis that the values for the difficulty, discrimination, and pseudo-chance parameters for individual ICCs are the same for the two groups. Items within the verbal and mathematics sections of the test were ranked according to their chi-

square values. The mean of the absolute difference between ICCs (Cook, Eignor, & Peterson; 1985) was then calculated for items with relatively small chi-square values. This new difference statistic was used because it, unlike the chi-square statistic, detects differences in ICCs when non-uniform differences are present. From those items with the smallest chi-square values, verbal and mathematics items with smaller mean absolute differences were considered equivalent and used as potential anchor items to equate the two language versions of the test. It should be noted that consideration was given to the language of origin, item type (e.g., antonyms, analogies) for verbal items and content area (e.g., algebra, geometry) for mathematics items when the final set of equating items was chosen.

The two examples presented here illustrate the use of three of the more popular methods of establishing translation equivalence. In both of these examples, a judgmental method (more specifically, method 1.B.1 - source language monolinguals check for errors) of establishing translation equivalence was used before applying a statistical method for the same purpose. That method 1.B.1 was used in these two examples is not unusual. Method 1.B.1 is by far the most common judgmental method of establishing translation equivalence in use today and is used almost routinely as a general check of translation equivalence.

The two examples also illustrated the use of the two more popular statistical methods of establishing translation equivalence. These include method 2.A.1 (bilinguals take source and target versions) in the first example and, in the second example, method 2.B.1 (source language monolinguals take source version and target language monolinguals take target version). The use of method 2.B.1 is, however, a more recent

trend due to the established feasibility of using item response models in conjunction with this method. The advantages of using item response models as a conditional statistical technique with this method were introduced in the previous section and will be discussed further in section 2.3.2.

2.3 The Use of Item Response Models in Establishing Translation Equivalence

2.3.1 Introduction

The discussion presented in section 2.2.4 highlighted the advantages of using method 2.A.2 (source language monolinguals take source version and target language monolinguals take target version) for establishing translation equivalence. The main advantage of this method is that translation equivalence results based on its use are more generalizable to the populations of interest (source and target language monolinguals) than with other methods of establishing translation equivalence. The main disadvantage of this method is that these results can be confounded with ability differences between the two samples of examinees. However, these ability differences can be controlled for by applying a conditional statistical technique when comparing examinee responses. Although a number of conditional statistical techniques are available for this purpose, the use of item response models is theoretically preferred when comparing groups of examinees who differ in ability (Ironson, 1983; Hambleton & Swaminathan, 1985). For this reason and additional reasons to be discussed in section 2.3.2, the focus of attention will now shift to the use of item response models in establishing translation equivalence. Section 2.3.2 will present the advantages of using item response models to establish translation

equivalence. Sections 2.3.3. and 2.3.4 will focus respectively on the preliminary considerations and steps in using item response models for this purpose.

The item response models discussed in sections 2.3.3 and 2.3.4 are those that are commonly used in practice for test development, test evaluation, and other testing applications. Two important points about these models are that they are designed for use with (a) unidimensional tests (that is, the test being used measures one dominant underlying trait) and (b) dichotomously scored test data. Item response models that do not require these restrictions have been developed; however, these models are relatively complicated and computer programs for estimating item and ability parameters from these models are not readily available. For these reasons, the discussions that follow will be based on the commonly used one-, two-, or three-parameter unidimensional logistic models.

2.3.2 Advantages of Using Item Response Models to Establish Translation Equivalence

The use of item response models has received much recent attention as a statistical technique for establishing translation equivalence (Candell & Hulin, 1987, 1987; Hulin, Drasgow, & Komocar, 1982; Irvine & Carroll, 1980; Hulin & Mayer, 1986; Poortinga, 1983; Simon, 1989; van der Flier, 1982; van de Vijver & Poortinga, 1991). The reason for this attention is that the framework of item response theory provides potential advantages over other conditional statistical techniques when establishing translation equivalence. These advantages can be obtained when an item response model provides a reasonable fit to the test data and include (1) item statistics (parameters) that are independent of the

specific sample of examinees used to calibrate the items; (2) examinee ability estimates that are independent of the specific choice of test items used from the calibrated item pool; and (3) examinee ability estimates of known precision. Of particular importance in a translation equivalence study is the first advantage - invariant item parameter estimates.

Invariant item parameter estimates are particularly useful in a translation equivalence study because they provide a strong basis for taking into account differences in examinees abilities when comparing item parameters across populations. Comparisons of item parameters across populations can be carried out by a number of different conditional statistical techniques (see section 2.2.4) other than the use of item response models. However, these alternative techniques can be problematic. For example, those methods based on the chi-square statistic are sensitive to sample size and the number of total score intervals used (Ironson, 1982). The Mantel-Haenszel statistic provides a close approximation to results obtained using the one-parameter logistic model but fails to flag items when non-uniform bias is present (Hambleton & Rogers, 1989). When it is possible to use them, item response models are generally preferred for identifying items that are functioning differently across populations because they (1) explicitly state the relationship between examinee ability and the probability of obtaining a correct response on an item and therefore are a more direct way of identifying differentially functioning items and (2) provide invariant parameter estimates (Ironson, 1983; Mellenbergh, 1983, 1989).

It should be noted that invariant examinee ability estimates are also of interest in the context of designing and using translated tests

for comparing examinees across populations. When using item response theory in a translation equivalence study, items that did not translate well (non-equivalent items) can be placed on the same ability (or difficulty) scale as those that did translate well (equivalent items). Hulin (1987) noted two benefits of using non-equivalent items when comparing examinees across populations. The first benefit is that instruments can be designed and administered that are potentially more meaningful to the populations of interest:

> The potential for producing equated scales containing mixtures of both emic[5] and etic items offers an additional advantage of IRT procedures in translation and cross-language research. Assuming there are a number of well-translated etic items and that the new emic items meet the assumption of IRT and reflect differences in the same unidimensional latent trait as the culturally general etic items, investigators can tailor scales to each culture by adding a number of emic items specific to each culture to the common set of culturally general etic items. This should increase the sensitivity and cultural relevance of the instrument for both cultures, yet retain the psychometrically required property of equated trait estimates. (Hulin, 1987)

If the items within an instrument are more meaningful to examinees within a population, it is likely that the instrument will also have greater reliability and validity within the population.

The second benefit of using non-equivalent items when comparing examinees across populations is that the precision of examinee ability estimates in each population is increased:

> The presence of many emic concepts in the source language of a particular scale would generate evidence of psychometrically non-equivalent items across the source and target language versions of the instrument. The nonequivalent items could be eliminated and conclusions about θ could be based on the items that were well translated and

[5]The term emic refers to terms or concepts that are specific to a population. Its counterpart, etic, refers to terms or concepts that are universal across populations.

> met the criterion of psychometric equivalence above. However, this involves eliminating the item from both versions of the questionnaire. If the translated item is nonequivalent in the target language but has a nonzero slope for the target language ICC, the item still provides information about θ in both cultures. The information about θ in both languages and cultures provided by the revised scale after eliminating all nonequivalent items would be less than if the entire scale consisting of the complete set of items were scored and used to estimate θ. Cross-cultural comparisons based on more information about θ in both cultures are more precise. (Hulin, 1987)

Both of these additional benefits of using non-equivalent items when comparing examinees across populations accrue from invariant examinee ability estimates that can be obtained within the framework of item response theory. Even though these additional benefits are not directly related to establishing translation equivalence (these benefits can only be obtained after completing a translation equivalence study), they offer further compelling reasons for using the framework of item response theory in comparing examinees across populations where differences in language or culture exist.

The advantages of using item response models over other conditional statistical techniques in establishing translation equivalence are gained at a cost. Aside from practical considerations such as the use of large sample sizes and relatively complex numerical procedures, restrictive assumptions about the test, its administration and the resulting scores must be made. These assumptions, which will be discussed in later sections, include (1) test unidimensionality, (2) non-speeded test administration, and (3) an adequate fit of resulting test scores to an item response model (Hambleton & Swaminathan, 1985). Each of these assumptions make it less likely that item response models can be used to establish translation equivalence. However, these assumptions can be checked and, when they are met, the advantages

provided by using item response models in cross population comparisons are both unique and extremely useful.

2.3.3 Preliminary Considerations to Using Item Response Models to Establish Translation Equivalence

When an item response model provides a reasonable fit to test data from the populations of interest, the benefits of using an item response model to establish translation equivalence described in the previous section can be obtained. However, consideration must be given to four factors before deciding to use item response models for this purpose. If any of these factors is considered a problem, item response models cannot be used in a translation equivalence study.

The first preliminary consideration is cost. Estimating the item and ability parameters associated with item response models generally requires the use of computer programs. These programs are relatively expensive to purchase (for example, the PC version of BILOG - an item and ability estimation program for the one-, two-, and three-parameter logistic models - costs approximately $300). Also, these programs are relatively expensive to run on mainframe systems. This is particularly true when the three-parameter model is used (estimating three item parameters uses a relatively large amount of computer time as compared to estimating one- or two-item parameters) or an item parameter is difficult to estimate (most notably; estimating the pseudo-chance (c) parameter when data from small numbers of low-ability examinees are available). Other programs may also be used for addressing model-data fit or for comparing item characteristic curves, thus adding to the cost of using item response models.

The second preliminary consideration is the availability of a minimum sample size. The minimum sample size recommended for use with the one-parameter logistic model is 200 examinees for a 20-item test (Wright & Stone, 1979). Since the one-parameter model requires the smallest number of examinees for accurate item and ability parameter estimates, a sample size of 200 examinees in each of the populations of interest is an appropriate minimum sample size for considering the use of item response models in establishing translation equivalence. Larger sample sizes are required when considering the use of the two- or three-parameter logistic models. Guidelines of minimum sample sizes required for using the different item response models are discussed in section 2.3.4.

The third preliminary consideration is the item scoring used in a test. As mentioned previously, the more commonly used item response models require dichotomously scored test data. In many instances, this is not an issue since dichotomous scoring with a variety of item formats is commonly used. For example, true-false, multiple choice, and sentence completion items are typically scored either right or wrong. However, it may be of interest to use polychotomous scoring with certain item formats. For example, multiple choice items may be scored by applying scoring weights to the different item options to obtain more information from each item. In this case, commonly used item response models that require dichotomously scored test data cannot be used.

The fourth preliminary consideration is the dimensionality of the tests being used. As mentioned previously, the more commonly used item response models are unidimensional models. The use of these models requires the assumption that examinee responses to all of the items in a

test can be attributed to one dominant underlying trait or ability. Unless the unidimensionality assumption can be met by examinee test data in the populations of interest, commonly used (i.e., unidimensional) item response models cannot be used to establish translation equivalence.

A number of methods can be used to check the assumption of unidimensionality in a set of test items. Hattie (1984; 1985) provides a thorough review of these methods of checking test dimensionality.

In summary, four factors should be considered before deciding to use item response models to establish translation equivalence. These factors are: (1) cost, (2), sample sizes, (3) item scoring, and (4) test dimensionality. If any of these four factors is considered a problem and steps cannot be taken to eliminate the problem, item response models should not be used to establish translation equivalence.

2.3.4 Steps in Using Item Response Models to Establish Translation Equivalence

This section provides an overview of the steps in using item response models to establish translation equivalence. These steps include: (1) model selection, (2) scaling of item and ability parameters, (3) comparisons of item characteristic curves (ICCs), and (4) evaluation of translation equivalence.

1. Selection of a Model. The first step in using an item response model to establish translation equivalence is deciding which model to use. As was discussed in section 2.3.1, the more commonly used item response models are the one-, two-, and three-parameter logistic models. These models should only be used with unidimensional test data that is dichotomously scored. Alternative models that can handle non-

dichotomously scored or multidimensional test data are not practical to use at this time for the reasons discussed in section 2.3.1. Therefore, the present discussion will be limited to choosing between the one-, two-, and three-parameter logistic models.

As was the case when deciding whether to use item response models to establish translation equivalence or not, practical considerations also play a role in deciding which model to use. The first practical consideration is the availability of sufficient samples of examinees. Estimates of the sample sizes needed for accurate estimation of parameters in item response models must be considered in light of several factors. These include (Hambleton, 1979): (a) test length (in general, shorter tests require larger sample sizes); (b) the parameter estimation method used (in general, Bayesian methods give more accurate parameter estimates with smaller sample sizes than maximum likelihood methods) and (c) the distribution of ability in the examinee samples (in general, larger sample sizes are required when homogeneous rather than heterogeneous samples are used).

However, some general guidelines regarding required sample sizes for using the different item response models are available. Hambleton (1979) provides a summary of the minimum test length and sample sizes required to obtain satisfactory ability and item parameter estimates using maximum likelihood estimation procedures. The following minimum test lengths and sample sizes were reported: 20 items and 200 examinees for the one-parameter logistic model; 30 items and 500 examinees for the two-parameter logistic model; and 60 items and 1,000 examinees for the three-parameter logistic model. These guidelines should be considered general rules-of-thumb for minimum test lengths and sample sizes. The

actual sample size required will depend on the three factors mentioned previously.

The second practical consideration in deciding which item response model to use is the nature of the examinee samples. Hambleton & Swaminathan (1985) point out that "Size of the sample is certainly important (in choosing a model) but so is the nature of the available sample. For example, if a three-parameter model is chosen and the sample is such that only a few examinees at the low ability level are available, then the chance level parameter cannot be estimated well. The three-parameter model should not be chosen in this case." The same authors also point out that "Alternatively, it may often be reasonable to choose a priori a constant value for the 'c' parameter." However, anticipating the nature of an examinee sample in a target population may be a difficult, if not impossible, task (an extreme example would be a cross-cultural study involving an isolated culture). In such cases, it may be wiser to forego the use of the three-parameter model than to venture a guess at a constant value for the "c" parameter. Instead, the one-or two-parameter models should be considered as alternatives.

The third practical consideration in deciding which item response model to use is the availability of computer programs to estimate ability and item parameters. A small number of computer programs are available for this purpose (a list of the more commonly used programs and their characteristics is provided in Appendix B). However, as can be seen from Appendix B, each computer program has characteristics which can limit its use in certain applications. The first potential limitation is the type of model for which the program can be used. For example, BICAL cannot be used with the two- or three-parameter logistic

model and therefore one of the other programs would be required to use these models. The second potential limitation is the estimation procedure used in the program. For example, if the user is interested in setting Bayesian priors to facilitate the estimation procedures, either BILOG or MicroCAT must be used. A third potential limitation is the computing environment an estimation program can be used in. All of the programs (with the exception of MicroCAT) will run on mainframe systems; however, only BILOG and MicroCAT are available for personal computers. Each of these limitations individually or taken together can influence the final choice of the item response model used in a translation equivalence study.

Further consideration of which item response model to use should depend on the characteristics of the test data. More specifically, the question to be asked is: How realistic are the assumptions of a model for test data from the populations of interest? These assumptions include (Hambleton & Swaminathan, 1985): (a) equal discrimination indices for the one-parameter model, (b) minimal guessing for the one- and two-parameter models, and (c) non-speeded test administrations for the one-, two-, and three-parameter models.

These assumptions can be checked for the populations of interest by a variety of methods. For example, the assumption of minimal guessing can be checked by examining the frequency of item options responded to by examinees. When each item option for a number of items is chosen with the same or approximately the same frequency, it is likely that examinees are guessing at the answers to those items. This and other checks on the characteristics of the test data can aid in deciding which item response model should be used.

The results obtained from checking test data for model assumptions must also be considered in light of model robustness. Model robustness refers to the extent to which the assumptions of a model can be violated and still lead to useful results. For example, an assumption when using the one-parameter model is that guessing is not a factor in examinee responses to test items. However, suppose the one-parameter model was used and examinees did guess at the answers to items. How useful would the results obtained from using the model be?

In one respect, it depends on the specific application the results will be used for. For example, if the purpose of using an item response model is to provide a bank of calibrated items for administering adaptive tests, invariant ability estimates are of particular importance. The robustness of a model with respect to this expected model feature becomes especially important. In contrast, if the purpose of using an item response model is to horizontally equate two versions of a test, invariant ability estimates are less important and therefore model robustness with respect to invariant ability is also less of a concern.

Unfortunately, the results of model robustness studies with respect to those applications which can effect translation equivalence studies have not provided clear guidelines on the extent of model robustness for these applications. Furthermore, guidelines in this area may be difficult to come by in general:

> There is some evidence that the models are robust to moderate departures from the assumptions, but the extent of 'model robustness' has not been fully established (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978), and it probably cannot be fully established. This follows because there are a myriad of ways in which model assumptions can be violated, and the seriousness of the violations depends on the nature of the examinee sample and the intended application. (Hambleton, 1979)

The problem of establishing model robustness to violations of model assumptions makes choosing a model based on the characteristics of the test data more difficult.

If the test data violate the assumptions (excluding the unidimensionality assumption which applies to either the one-, two-, or three-parameter model) of a particular model, a researcher has the option of (aside from not using an item response model) using a particular model and then assessing the degree of model-data misfit. If the fit of the test data to the model is reasonable, the model can be used confidently to establish the degree of translation equivalence. Model-data fit will be discussed next.

The next step after initially choosing a model and estimating item and ability parameters is to assess model-data fit. This step is important because the advantages of using an item response model can only be obtained when a model provides a reasonable fit to the test data.

Evidence concerning model-data fit can be gathered by applying a variety of approaches. These approaches for addressing model-data fit can be placed into two general categories (Hambleton & Swaminathan, 1985). The first general category is checking expected model features and includes approaches for checking the invariance of item and ability parameter estimates. For example, the invariance of ability estimates can be checked by comparing ability estimates from two or more item subsets from the item pool. The item subsets chosen for comparison are often based on differences in item difficulties (e.g., relatively easy vs. relatively hard) or content categories. In the case of a translation equivalence study, these or other item subsets of interest

(for example, items that were more difficult to translate vs. items that were not) can be used.

The second general category of approaches for addressing model-data fit is checking model predictions of actual or computer-simulated test results. An example of an approach for checking model predictions with actual test results is the use of residual analysis. This approach makes use of model predictions of the item performance of examinees and compares them to the actual results for each item obtained for various ability groups within the sample. The resulting residuals can be standardized to allow for an interpretation of the model-data fit in terms of the standard normal distribution.

The two examples presented here are illustrative of the general approaches for addressing model-data fit. Further details on these and other approaches for addressing model-data fit are available in Hambleton and Swaminathan (1985) and Hambleton and Murray (1983).

2. Scaling of Item Parameter and Ability Estimates. Assuming that a reasonable fit between an item response model and the test data for examinees in each population have been obtained, the next step in establishing translation equivalence is to place the item parameters obtained in each population onto a common ability (or difficulty) scale. Item parameter and ability estimates obtained in each population are each defined on different ability (or difficulty) scales because of differences in the ability levels of the examinee samples (it is highly unlikely that the ability levels of examinee samples across populations would be equal). Test or item characteristic curves cannot be compared across populations until a common metric has been established.

There are three main designs for scaling two or more versions of a test. These include the (a) single-group, (b) random-groups, and (c) anchor test designs (Cook & Eignor, 1983). The single-group design cannot be used when establishing translation equivalence since more than a single group of examinees is involved. Also, the random-groups design cannot be used in this context since two distinct populations are involved and, therefore, two equivalent groups cannot be selected. The only viable scaling design that can be used in a translation equivalence study is the anchor test design. This design makes use of "common items" to anchor each version of a test to a common scale.

The procedure for scaling the source and translated versions of a test with the use of anchor items includes three steps:

(1) initially scaling the item or ability parameters obtained from the source and translated versions of a test;

(2) comparing ICCs for corresponding items from the source and target versions of the test; those items with the same ICCs are designated as anchor items; and

(3) scaling the source and translated versions of the tests using the designated anchor items.

Some of the details associated with each of these steps will be discussed next.

Step 1 - The initial scaling of the item and ability parameters for test data obtained from source and target language examinees can be accomplished using a variety of methods. These methods include: (a) concurrent calibration, (b) the equated bs method, and (c) the characteristic curve method (Peterson, Kolen, & Hoover, 1989). A discussion of these scaling methods is presented in section 2.4.2.

The anchor items used with these scaling methods when establishing translation equivalence are typically all of the items in a test. The assumption is initially made that corresponding items in each version of the test are equivalent. If this initial assumption was not made, a basis for scaling the two versions of the test would not exist.

An alternative method would be to develop anchor items specifically for this purpose. Source language items could be developed that might be easily translated into items that are suitable for target language examinees. These items could be used as anchor items if they measure the same underlying trait as the remaining items in a test. The use of this method has not been mentioned in the test translation literature.

The use of test items developed specifically for use as anchor items might suggest that steps 1 and 2 of the scaling process are not necessary. If anchor items are available, why not skip to step 3 and simply scale the source and target versions of a test? The problem is that it is not really known whether these specifically designed anchor items are in fact equivalent across language versions of a test. Judgments about their equivalence may have been made, but their statistical equivalence has not been established. Therefore, a more viable alternative is to incorporate the specifically designed anchor items into a translation equivalence study and proceed with steps 1 to 3 of the scaling procedure as it is usually done. If the specifically designed anchor items are actually equivalent, they will likely emerge as the anchor items used to scale the source and language versions of a test.

Step 2 - The next step in scaling the source and target versions of a test is to compare the ICCs for corresponding items from each version of the test. ICCs can be compared by several different methods (Ironson, 1983; Hambleton & Swaminathan, 1985; Mellenbergh, 1989). One method is to calculate the area between ICCs. Either a direct measure of this area or, to take into account possible non-uniform differences in ICCs, absolute values of squared differences of the area between ICCs can be computed.

A second method of comparing ICCs is to compare the parameters for ICCs. In the most general case (comparing ICCs from a three-parameter model), a Chi-Square statistic can be used to test the null hypothesis of equal item parameters ($H_o: a_{1i}=a_{2i}$, $b_{1i}=b_{2i}$ and $c_{1i}=c_{2i}$). For the two-parameter model, only the a's and b's would be compared while only the b's would be compared when a one-parameter model is used.

A problem when comparing item parameters for the three-parameter model may arise because of poor estimation of the c-parameter in one or both of the populations being compared. This can be caused by a lack of low ability examinees to properly estimate the c's. Therefore, the estimation procedure proposed by Lord (1980, chapter 14) should be used before step 1 (initial scaling of item and ability parameters) when using a three-parameter model if ICCs are to be compared using item parameters and the value of the c's are not fixed. The steps of this procedure are:

1. Combine the test data from source and target language examinees and estimate the item and ability parameters standardizing on the b's.

2. Holding the c values obtained in the previous step fixed, estimate the a, b, and ability parameters for each language group separately.

This procedure results in c values that are the same for corresponding items in the source and target versions of the test. Therefore, only the a and b parameters are compared across language groups when this estimation procedure is used.

The third method of comparing ICCs is to compare fit statistics for the ICCs. The rationale behind this method of comparing ICCs is that an item that shows no difference between ICCs should either fit or misfit a particular model in the same way for each of the populations being compared. The usefulness of this method has not been established and it is relatively unpopular compared with the first two methods of comparing ICCs.

For detailed discussions of these methods and their relative merits, see Berk (1982), Ironson (1983), or Hambleton and Swaminathan (1985).

Once ICCs from the source and target versions of a test have been compared by one of the three methods outlined above, those items with the same ICCs can be used as anchor items to scale the two versions of the test. However, a problem with using these items as anchor items may exist. This problem was pointed out by Lord (1980) in the context of item bias studies. If many of the items are not statistically equivalent, then the set of items being compared may not measure a unidimensional trait. Consequently, the ICCs for the source and target examinees may not be directly comparable.

A potential solution to this problem is given in Lord (1980, chapter 14). This solution purifies the criterion used to scale the two versions of the test and includes the following steps:

1. Analyze the source and target versions of the test using steps 1 (initial scaling) and 2 (comparison of ICCs) discussed previously.
2. Remove all test items that have significantly different ICCs.
3. Combine the test data from source and target language examinees and estimate the ability parameter for each examinee.
4. Replace the items removed in step 2.
5. Holding the ability parameter estimated for each examinee in step 3 fixed, estimate the a and b parameters for each item in each language version of the test.

The resulting ability (or difficulty) scale is now more likely to be based on a set of unidimensional items. ICCs for corresponding items across examinee groups are again compared. Comparisons of ICCs based on this "purified" scale are potentially more meaningful.

Step 3 - Those items identified as equivalent items in step 2 are then used as anchor items to scale the source and translated version of a test. Any of the methods of scaling item and ability parameters mentioned previously can be used to place these parameters on a common scale. Once these parameters are on the same scale, the source and target versions of a test are equated.

In summary, the steps for scaling item and ability parameter estimates for test data from source and target language examinees are:

1. Obtain separate item and ability parameter estimates for source and target language examinees. If a three-parameter model is being used, and the comparison of ICCs is to be based on a

comparison of item parameters, use the parameter estimation procedure proposed by Lord (1980, chapter 14) to reduce the problem of inaccurately estimated c's.

2. Scale the item and ability parameter estimates using one of the available scaling methods.
3. Compare ICCs for corresponding items from the source and target version of the test.
4. Temporarily remove items with significantly different ICCs from the item pool.
5. Combine the test data from source and target language examinees and estimate each examinee's ability.
6. Replace the items that were removed in step 4.
7. Holding the ability parameter (estimated in step 5) for each examinee fixed, estimate the a and b parameters for each item in both the source and target language versions of a test.
8. Scale these item parameter estimates using those items identified as equivalent in step 2 as anchor items.
9. Compare ICCs for corresponding items across each language version of the test.
10. Using items with the same ICCs in each group as anchor items, scale the item parameters from the source and translated versions of a test using one of the available scaling methods.

3. Comparison of ICCs. After the iterative procedure mentioned previously for scaling the source and translated versions of a test has been completed, ICCs for corresponding items in each version of the test can be compared to determine the extent of translation equivalence for

individual items. Any of the three methods of comparing ICCs mentioned previously can be used at this step.

4. Evaluation of Translation Equivalence. Once ICCs for corresponding items in the source and target version of a test have been compared, final decisions about the extent of translation equivalence for individual items must be made.

First, a decision about what degree of differences in source and target version ICCs will constitute translation non-equivalence must be made. When ICCs are compared using any of the comparison methods, some differences in the ICCs are to be expected even for equivalent items because of errors in estimating the item parameters in each sample. Therefore, one decision is how much difference in the ICCs should be attributed to these errors.

A potential solution to this problem has been suggested by Rogers and Hambleton (1989). The authors suggest evaluating the difference between ICC's in reference to a baseline generated through computer simulation. By generating a set of data that reflects the test data of interest but with no bias present, comparisons of the ICC's for the same items when no bias is present is possible. From these comparisons of ICCs, a sampling distribution of a bias index under the condition of no bias can be generated and a realistic cut-off value chosen to interpret the bias index for the data of interest.

It may also be unrealistic to expect ICCs from the source and target versions of a test to be exactly the same even if errors in estimating item parameters could be eliminated. Therefore, a decision about how much difference in ICCs is acceptable (if any) before items are considered non-equivalent might also need to be made. In one sense,

when statistical tests of item parameters are used to compare ICCs, this decision is somewhat easier since differences between the corresponding item parameters being compared at a specific significance level signify lack of translation equivalence. However, the significance level for these tests must be decided upon and consideration must be given to the sample sizes used. Both of these factors can influence the results of statistical tests and therefore the results of a translation equivalence study.

Second, a decision about what differences in ICCs really mean must be made. Consider the following example. Suppose corresponding ICCs for an item from the source and target versions of a test are considered different. Two possible explanations for this difference exist. One possible explanation is that the two populations being compared differ on what the item is measuring. In this case, the item is correctly measuring a different trait of interest in each population. The second possible explanation is that the item did not translate well and therefore it is not measuring the same trait in the populations being compared. Careful consideration must be given to each of these possible explanations when differences in ICCs are evident before making a decision about what the differences actually mean. This problem points out the importance of obtaining evidence of construct validity for each of the items in each of the tests being used.

In summary, the following steps are required when using an item response model to establish translation equivalence:

1. Selection of an item response model.
2. Scaling of item parameter and ability estimates.
3. Comparison of ICCs, and

4. Evaluation of translation equivalence.

The amount of effort put into each one of these steps will ultimately determine the validity of a translation equivalence study.

2.4 Test Scaling Through Item Response Theory

2.4.1 Introduction

Item and ability parameter estimates obtained from different groups of examinees must be placed on a common scale before ICCs can be compared across groups. As noted in section 2.3.4, the only appropriate scaling design that can be used when attempting to establish the translation equivalence of test items is the use of anchor items. The different methods of scaling item and ability parameter estimates using anchor items can be classified in two categories depending on whether the parameter estimates are calibrated simultaneously or separately. These methods will be briefly discussed in the following section. The scaling method used in this study will be discussed in detail in section 3.5.

2.4.2 Methods of Scaling Parameter Estimates

Scaling Method used with Simultaneous Parameter Estimation

An option available to users of the LOGIST parameter estimation program (Wood, Wingersky, & Lord, 1976) is concurrent calibration. In concurrent calibration, item and ability parameter estimates for each examinee group are estimated simultaneously by 1) coding the unique items for test X as not reached by examinees who took test Y, 2) coding the unique items for test Y as not reached by examinees who took test X, and 3) using LOGIST to simultaneously estimate the parameters. Provided access to the LOGIST program is available for use, concurrent calibration is a convenient method of scaling item and ability parameter

estimates since the calibration and scaling procedures are performed simultaneously.

Scaling Methods used with Separate Parameter Estimation

It is also possible to obtain item and ability parameter estimates for each examinee group separately. When separate estimation of parameters is used, the scale on which ability is defined is somewhat arbitrary (Hambleton & Swaminathan, 1985). For example, when LOGIST is used, $\beta = 0$ and $\sigma = 1$ for each set of parameter estimates. Since the ability estimates in each examinee group will most likely be different, the ability scale and choice of origin for parameter estimates calibrated separately are not comparable. However, ability estimates from each group of examinees are linearly related. This linear relationship is given by $\theta_y = \alpha\theta_x + \beta$ for test versions x and y. This scaling or equating line should, in theory, be a straight line. However, because of parameter estimation errors, the actual point estimates of the θs will be scattered about this "linear" scaling line.

A number of different scaling methods can be used to place these linearly related ability estimates on a common scale when ability estimates are obtained in separate calibration runs. One of the easiest scaling methods to apply is the mean and sigma method (one of the "equated b's" methods). Since difficulty estimates are on the same scale as ability estimates (θs), the linear relationship for ability estimates from two groups of examinees given by $\theta_y = \alpha\theta_x + \beta$ can also be applied to difficulty estimates from items on the two test versions to be scaled. Moreover, mean difficulty estimates can be used instead of individual item difficulty estimates to obtain the scaling constants α

and β. Thus, the following relationships provide a basis for parameter scaling using the mean and sigma method (Hambleton & Swaminathan,1985):

$$\alpha = \frac{S_y}{S_x} \qquad [2.4.1]$$

and

$$\beta = \bar{b}_y - \alpha\bar{b}_x \qquad [2.4.2]$$

The steps for implementing the mean and sigma method are outlined in Hambleton and Swaminathan (1985) and Crocker and Algina (1986). The advantage of this method is that it is relatively easy to implement. The disadvantages of this method are that it does not take into account 1) the varying accuracy of item and ability parameter estimates and 2) outlying point estimates of ability or item difficulty on the calculations of the scaling coefficients α and β (Hambleton & Swaminathan, 1985). To reduce the problems associated with the first disadvantage of the mean and sigma method, the robust mean and sigma method was introduced by Linn, Levine, Hastings, and Wardrop (1981). This method uses weights based on standard errors to reduce the influence of poorly estimated parameters on the calculations of the scaling coefficients. Stocking and lord (1983) added additional steps to the robust mean and sigma method to reduce the problems associated with the second disadvantage of the mean and sigma method. This robust iterative weighted mean and sigma method takes into account the perpendicular distance of ability or item difficulty point estimates from the scaling line calculated using the robust mean and sigma method. The scaling line is iteratively adjusted by reducing the influence of outlying point estimates on the scaling coefficients.

A disadvantage of each of these mean and sigma methods is that they do not make use of all of the item and ability parameter information that is available when scaling tests. The characteristic curve method (Stocking & Lord, 1983) was developed to alleviate this disadvantage of other scaling methods. This method minimizes the mean squared difference between the true scores for each examinee j on test x (ξ_j) and the transformed true score on test y(ξ_j^*). Studies comparing different scaling methods have generally concluded that the characteristic curve method provides relatively accurate scaling results compared to other scaling methods (Stocking & Lord, 1983; Johanson, 1987; Wingersky, Cook, & Eignor, 1987).

Divgi (1985) has, however, noted two disadvantages with the characteristic curve transformation scaling procedure. First, the method is relatively complex and therefore requires more computer time to implement than other scaling methods. Secondly, the procedure does not take into account the standard errors of the parameter estimates. Divgi (1985) proposed the minimum chi-square method in order to overcome these disadvantages. The minimum chi-square method incorporates the covariance matrix of sampling errors when minimizing the difference between the discrimination (a) and difficulty (b) parameters for test x and the transformed a and b parameters for test y. This method is potentially useful but to date has not received much attention in the test scaling literature.

2.4.3 Anchor Test Length and the Scaling of Parameter Estimates

Anchor items are required to place item and ability parameters estimated from two different examinee samples on the same scale. Once placed on the same scale, these parameters can be compared to establish

translation equivalence. In the context of a translation equivalence study, it is unlikely that the different language versions of a test will have similar difficulties and/or that examinees in the populations being compared will have similar mean abilities. Therefore, the studies reviewed here are concerned with vertical rather than horizontal scaling.

A number of studies have been conducted on the anchor test length required for adequately scaling item and ability parameter estimates. McKinley and Reckase (1981) investigated the number of anchor items required for scaling parameter estimates when developing a large calibrated item pool. They used real achievement test data that covered a variety of subjects areas. The authors concluded that with concurrent calibration, 25 anchor items provided better scaling results than 15 anchor items, but that 15 anchor items provided adequate scaling results. However, Wingersky and Lord (1984) point out that these results should be regarded as suspect since "their data clearly violated the unidimensionality assumption."

Vale, Maurelli, Gialluca, Weiss, and Ree (1981) investigated the number of anchor items required for test scaling when the shape of the information curve for the anchor test varied. Simulated data for 4,000 examinees and anchor test lengths of 5, 15, and 25 items were used. Linear scaling of the ability estimates for anchor items was used to place the unique items onto a common scale. Item parameter estimates obtained from the three-parameter logistic model for the unique items were compared with their true values (known because the data was simulated) to evaluate the adequacy of the scaling. Vale et al. concluded that 5 to 25 anchor items provided adequate scaling and that

anchor tests with peaked test information curves gave poorer scaling results than those with normal or rectangular shaped information curves.

Raju, Edwards, and Osberg (1983) investigated the number of anchor items required for vertical scaling using real data. Both the one- and three-parameter logistic models were used. Item parameter estimates were estimated in separate calibration runs and scaled using the mean and sigma method. For both the one- and three-parameter logistic models, the authors concluded that 6 to 8 anchor items performed almost as well as 18 to 24 anchor items. The three-parameter model provided better overall results than the one-parameter model.

Wingersky and Lord (1984) investigated the number of anchor items required for test scaling using concurrent calibration. Anchor test sizes of 2, 25, and 50 items were used. Data for this study was obtained from examinee responses to two versions of the mathematics section of the SAT (descriptive statistics for these tests were not provided). The authors concluded that 2 good anchor items (items with low standard errors for the item parameters) provided similar scaling results to those obtained using 25 or 50 anchor items. However, the results from this portion of their study may have limited generalizability to vertical scaling situations since the difficulties of two versions of the SAT and/or the mean ability of the examinee samples were most likely not substantially different.

Vale (1986) investigated the number of anchor items required for test scaling when different test lengths and scaling designs were used. These scaling designs included the equivalent groups, anchor test (equated b's method) and interlaced designs. Simulated data for 750 to 1,000 examinees and four test lengths ranging from 31 to 40 items were

used. The number of anchor items ranged from 2 for the 31-item test to 20 for the 40-item test. For non-equivalent groups, Vale concluded that the interlaced scaling design worked best and that, even with the non-interlaced scaling designs, as few as 2 anchor items were required for adequate test scaling. Vale noted, however, that the number of anchor items used was confounded with test length and that the dimensionality of the test data was not considered, thus limiting the generalizability of these results.

Wingersky, Cook, and Eignor (1987) investigated the number of anchor items required for true score equating when the characteristics of the anchor items were systematically varied using simulated data. The characteristics of the anchor items studied included the standard error of the item parameter estimates, the shape of the ability distribution for the groups used to estimate the item parameters (uniform and peaked), model data fit for two of the anchor items, and item bias for two of the anchor items. The effects of these anchor item characteristics were investigated using two scaling methods: concurrent calibration and the characteristic curve method. The three-parameter logistic model was used throughout this study with sample sizes of 2,500 examinees used for each set of parameter estimates. The simulated data used in this study was generated to reflect typical item characteristics for the verbal section of the Scholastic Aptitude Test. The anchor test lengths used in this study were 10, 20, and 40 items.

Wingersky et al. concluded generally that 20 to 40 anchor items provided reasonable equating results. More specifically, the following conclusions were drawn from the study:

A. The results concerning the standard error of anchor item parameter estimates were counter-intuitive. Anchor tests consisting of anchor items with small standard errors generally provided less stable equating results than the same length anchor tests consisting of anchor items with average (typical for SAT-V) standard errors. The authors concluded that a possible explanation for these results was that the anchor items with average standard errors were more parallel in content and difficulty for the tests being scaled. They suggested that the relative efficiency of the anchor items (with respect to the total test) rather than standard errors of the item parameter estimates may be a preferred method of determining the quality of the anchor items.

B. Across the different anchor test lengths and scaling designs, better equating results were obtained when the distribution of examinee ability was uniform rather than peaked. Thus, when the distribution of examinee ability is peaked, a longer anchor test is required than when the distribution of examinee ability is uniform.

C. Use of two anchor items that were poorly fit by a three-parameter model did not significantly affect the equating results regardless of the number of anchor items used. The authors noted that this result may not be generalizable to situations where the number of misfitting items and the degree of misfit are different than those typically observed for SAT-V anchor items.

D. The use of two anchor items that functioned differently across the examinee groups profoundly affected the equating results

regardless of the length of the anchor test used. However, the equating results were most profoundly affected when (1) shorter anchor tests were used, and (2) the equating was based on the characteristic curve transformation method.

Johanson (1987) investigated the anchor test length required for adequate scaling with various scaling methods and percentage of ability overlap for examinee calibration samples. The scaling methods used included concurrent calibration, characteristic curve, mean and sigma, orthogonal least squares and ordinary least squares. Anchor test lengths of 4, 7, 13 and 25 items were used with examinee ability overlaps of 10%, 30% and 50%. Data for this study was simulated using examinee sample sizes of 500 in each group. The following conclusions were drawn from the study:

1. Across several combinations of scaling methods and mean group examinee ability differences, anchor test lengths as small as 4 items provided adequate scaling results.
2. The most accurate scaling results across all combinations of anchor test length and examinee ability overlap were obtained using the characteristic curve method.
3. Small examinee ability overlap most affected the scaling results when concurrent calibration was used and least affected the scaling results when the characteristic curve method was used.

Klein and Jarjoura (1985) investigated the effects on test scaling of using a longer anchor test to compensate for poor content representativeness in the anchor test. The tests used in this study were three versions of a 250-item, multiple-choice test covering six distinct content areas. The mean test length of the representative

anchor tests was 60 items while the non-representative anchor tests had a mean test length of 103 items. Both the equally reliable Levine scaling and Tucker linear (non-IRT) scaling procedures were used. The authors concluded that the use of an anchor test with poor content representativeness can adversely affect scaling results and that a substantial increase in anchor test length did not compensate for poor content representativeness of an anchor test.

The results of the seven studies where length of anchor test was included as an independent variable varied substantially. Both Wingersky and Lord (1984) and Vale (1986) concluded that adequate scaling results are possible with as few as 2 anchor items. In contrast, Wingersky et al. (1987) concluded that at least 20 anchor items were required and Reckase (1981) concluded that at least 15 anchor were required. The four remaining studies reviewed here concluded that the minimum anchor test length should be between these two extremes.

A number of factors may be responsible for the varied results obtained in these studies. For example, Cook and Eignor (1989) have noted that the sample sizes used in many of these length of anchor test studies were substantially different. Different calibration sample sizes can influence the accuracy of ability and item parameter estimation and therefore the accuracy of scaling results. A further possibility is that the degree of vertical scaling has been substantially different across a number of these studies. This situation may have resulted from using examinee samples with varying ability overlap or tests that vary in the degree of difficulty differences. The results of Johanson (1987) indicate that differences in the overlap of examinee ability can have a profound affect on the accuracy of scaling results.

Several other potential reasons for the varied results from these length of anchor test studies include differences in the scaling methods used, the dimensionality of the test data, model-data fit, and the methods used to evaluate the scaling results.

Other possible reasons for the varied results from these length of anchor test studies may be related to the nature of the anchor test used. For example, the results of Klein and Jarjoura (1985) indicate that differences in the content representativeness of the anchor test can affect the accuracy of scaling results. A further possibility is that the parameters of the anchor test items used in many of these studies may have varied in representativeness of the remaining test items. Statistical representativeness of the anchor test items may influence the anchor test length required to provide adequate scaling results.

In summary, a number of factors may have influenced the results of the length of anchor test studies reviewed here. Since a number of these factors exist, it is difficult to pinpoint the reasons for the difference in the results of these studies. Further understanding of the length of anchor test problem can only be obtained from additional studies that investigate the influence of these factors on the length of anchor test necessary to provide adequate scaling results.

CHAPTER 3

METHODS OF INVESTIGATION

3.1 Introduction

In this chapter, the procedures that were used in carrying out this study will be presented. This chapter is divided into six sections. Section 3.2 contains an overview of the study. Section 3.3 contains a description of the data used in this study. The procedures used in the study are listed in section 3.4. The scaling method used is described in section 3.5. Lastly, section 3.6 contains a description of the procedure used in evaluating the results from this study.

3.2 Overview of the Study

The purpose of this study was to investigate the anchor test length required to accurately scale parameter estimates obtained in two populations under a variety of conditions. More specifically, this study attempted to answer the following questions:

1. How do differences in calibration sample size affect the anchor test length required to provide reasonably accurate scaling results?
2. How do differences in the mean ability of examinee groups affect the anchor test length required to provide reasonably accurate scaling results?
3. How does the interaction of these two factors affect the anchor test length required to provide reasonably accurate scaling results?

And, finally,

4. What anchor test length will provide reasonably accurate IRT scaling results?

3.3 Description of the Data

In order to investigate the effects of calibration sample size and the overlap in the ability distributions of the populations being compared on the anchor test length required for accurately scaling parameter estimates, it was necessary to (1) know the true scaling constants and (2) be able to manipulate the variables being studied. Simulated data provided a means for accomplishing these goals. Examinee data generated through computer simulation procedures provided known item and ability parameters and, therefore, the scaling constants required to place these parameters on the same scale were also known. Deviations from these true scaling constants can come from two possible sources. First, errors due to the scaling method used are to be expected since no scaling method provides completely accurate results. Second, and more germane to the purpose of this study, scaling errors can result from the influence of the variables of interest on the scaling procedure. Simulated data provides a practical means of investigating this second source of scaling error since the variables of interest were readily manipulated.

An additional advantage of using simulated data was that extraneous variables that can be confounded with the variables of interest can be controlled. For example, when working within the framework of commonly used item response models, the multidimensionality of the test data being used is often an issue. When using simulated data, the potentially confounding affects of using multidimensional test data with a unidimensional item response model can be controlled for. However,

controlling extraneous variables can also reduce the generalizability of results based on simulated data compared to those based on "real" data. Therefore, simulated data should reflect the characteristics of "real" data as closely as possible while still allowing the variables of interest to be studied.

The data used in this study were generated using the computer program DATAGEN (Hambleton & Rovinelli, 1973). This program generates data sets based on user defined specifications. Using this program, data sets based on specific item parameters, ability distributions and other relevant characteristics can be generated.

All together, 32 data sets were generated for this study. These 32 data sets correspond to the cross between two examinee sample sizes (N=300 and 600 for each population), two levels of examinee ability overlap (50% and 80%), and four anchor test lengths (n=5, 10, 15 and 20). Each of these data sets represents examinee responses from two populations of examinees to two different language versions of a test. The examinee populations are designated group A (the source examinee sample) and group B (the target examinee sample). Group A was the lower ability sample and group B was the higher ability sample. The test taken by each group was designated test X and test Y respectively.

The two sample sizes used are both smaller than the minimum recommended sample size (N=1000) for parameter estimation using the three parameter logistic model with a 60-item test (Hambleton, 1979). The reason for using small sample sizes is that large examinee samples are typically not available when conducting translation equivalence studies. To reduce the problem of using relatively small sample sizes, a modified three-parameter logistic model with fixed pseudo-chance

parameters was used in this study. Fixing the pseudo-chance parameter at a specific value reduces the sample size required to accurately estimate item and ability parameters.

The two levels of ability overlap used represent a wide range of differences in the mean ability of the examinee groups being compared. Since each ability distribution will be normally distributed with σ=1, the 50% ability overlap corresponds to a mean ability difference of 1.35 and the 80% ability overlap corresponds to a mean ability difference of 0.51. The 50% ability overlap represents an extreme vertical scaling situation while the 80% ability overlap represents a less extreme vertical scaling situations.

The anchor test lengths used represent the range of anchor test lengths reported in the test scaling literature as necessary to provide accurate scaling results.

The examinee data for the unique (non-anchor) items (n=50) were generated under the following conditions:

1. Each examinee ability distribution was normally distributed with a standard deviation of 1.0. The mean of each ability distribution corresponded to the percentage of overlap in each of the data set. These means are -0.675 and 0.675 for the 50% ability overlap sample and -0.255 and 0.255 for the 80% ability overlap samples.
2. The mean item difficulty for each test was set to the appropriate group mean ability. Thus, the mean difficulty for test X was set to the mean ability of group A which varied depending on the ability overlap of the examinee samples. All item difficulties were uniformly distributed with a range of 1.5.

3. The mean item discrimination for each of the six tests was 1.0. All item discriminations were uniformly distributed with a range of 0.8.

4. All pseudo-chance values were set to 0.2.

These conditions are summarized in Tables 3 and 4.

The examinee data for the anchor items was generated under the same conditions as for the unique (non-anchor) items. Because this examinee data was randomly generated, the anchor test item parameters can be considered representative of those for the unique items, particularly for the longer anchor tests. For the shorter anchor tests, the anchor item parameters may not be as representative of the total test, but this potential problem is reduced by using replications of each data set. All data sets were replicated in order to obtain replications of the scaling results. The use of replications reduces the probability of obtaining inaccurate results due to chance fluctuations in parameter estimation.

3.4 Procedures

In this section, the procedures used in implementing this study are outlined. A step by step listing of these procedures follows.

Step 1 - Obtain item and ability parameter estimates for each of the 32 data sets described in the previous section (five replications were used for the data sets based on a sample size of 300 and 3 replications were used for the data sets based on a sample size of 600). These parameter estimates were obtained through the LOGIST parameter estimation program (Wood, Wingersky, & Lord, 1976) using a modified three-parameter logistic model (i.e., the three-parameter logistic model

Table 3

Means and Standard Deviations of the Ability Distributions for Groups A and B

Ability Overlap	Group A	Group B
50%	-0.675	0.675
	(1.0)	(1.0)
80%	-0.255	0.255
	(1.0)	(1.0)

All values in parenthesis are standard deviations.

Table 4

Means and Ranges of Item Difficulty, Discrimination and Pseudo-chance Parameters for Tests X and Y

Ability Overlap	Difficulty		Discrimination		Pseudo-Chance	
	Test X	Test Y	Test X	Test Y	Test X	Test Y
50%	-0.675	0.675	1.0	1.0	0.2	0.2
	(1.5)	(1.5)	(0.8)	(0.8)	(0.0)	(0.0)
80%	-0.255	0.255	1.0	1.0	0.2	0.2
	(1.5)	(1.5)	(0.8)	(0.8)	(0.0)	(0.0)

All values in parenthesis are ranges.

with the pseudo-chance parameters fixed at a specific value). The three parameter logistic model was used because it has been recommended over the one or two parameter logistic models for vertical scaling (Skaggs & Lissitz, 1986). The reason for this may be that for difficult tests (where a fair amount of guessing may occur), the three parameter model

provides a better fit to the test data than either the one or two parameter models.

Step 2 - Obtain the true scores for each set of examinees on the anchor items in their respective data sets. These true scores were obtained through a characteristic curve scaling program written in FORTRAN 5. A portion of this program calculates test characteristic curves from which the true scores for a set of examinees can be derived.

Step 3 - Obtain the scaling coefficients α and β using the characteristic curve scaling method (described in section 3.5) with the true scores for the anchor items obtained in step 2. The characteristic curve scaling program mentioned in step 2 was used to obtain the scaling coefficients.

Step 4 - Evaluate the accuracy of the test scalings for the true scores on the unique items for tests X and Y across the anchor test lengths for the various combinations of sample size and ability distribution overlap. The method used in evaluating the scaling results will be described in section 3.6. The scaling results were averaged across replications using a second computer program also written in FORTRAN 5.

Throughout these procedures, steps were taken to insure that poorly estimated item and ability parameters would not effect the results of this study. First, trial runs of LOGIST with the N=300 and 50% examinee ability overlap samples converged in 18 stages or less, indicating that, for the most difficult data sets to obtain convergence, obtaining convergence would not be a problem. As a check that LOGIST did not have difficulty converging for all of the data sets, it was set to run at a maximum of twenty stages. Convergence was obtained for all of the data

sets with this maximum stage setting. Second, in the computer programs mentioned under steps 2 and 3, unique test items that had absolute b values greater than 4.0 were removed from the analysis. This was done to eliminate the effect of poorly estimated b values on the scaling results. If a unique test item was removed from the analysis, the true scores for examinees were based on the number of remaining test items. If even one anchor item had an absolute b value greater than 4.0, the parameter estimates for that data set were not used and LOGIST was rerun for that data set. The number of items with poorly estimated b values was monitored throughout this analysis.

3.5 Characteristic Curve Scaling Method

In an effort to make use of all of the available item and ability parameter information when scaling tests, Stocking and Lord (1983) introduced the characteristic curve method. This scaling method minimizes the mean squared difference between the true score for each examinee j on the anchor items in test x (ξ_j) and the transformed true score on the anchor items in test y (ξ^*_j). ξ_j and ξ^*_j are related by

$$b^*_{yi} = \alpha b_{yi} + \beta \qquad [3.5.1]$$

$$a^*_{yi} = a_{yi}/\alpha \qquad [3.5.2]$$

where b_i is the estimated difficulty for item i and a_i is the estimated discrimination for item i. The designation * indicates that the parameter is expressed on the same scale as test x. The constants α and β are chosen to minimize the difference between ξ_j and ξ^*_j. This is accomplished by minimizing the function

$$F = \sum_{j=1}^{n} N^{-1} (\xi_j - \xi^*_j)^2 \qquad [3.5.3]$$

with respect to α and β where n is the number of examinees.

The function F is minimized by setting the partial derivatives of F with respect to α and β equal to zero.

$$\frac{\partial F}{\partial \alpha} = \frac{-2}{N} \sum_{j=1}^{n} (\xi_j - \xi^*_j) \frac{\partial \xi^*_j}{\partial \alpha} = 0 \qquad [3.5.4]$$

$$\frac{\partial F}{\partial \beta} = \frac{-2}{N} \sum_{j=1}^{n} (\xi_j - \xi^*_j) \frac{\partial \xi^*_j}{\partial \beta} = 0 \qquad [3.5.5]$$

Since $\xi^*_j = \Sigma P^*_i(\theta_j)$ where $P^*_i(\theta_j) = P_i(\theta_j, a^*_i, b^*_i, c_i)$ for an examinee j with ability θ_j, the partial derivatives of ξ^*_j with respect to α and β in equations 3.5.4 and 3.5.5 can be solved for in terms of $P^*_i(\theta_j)$,

$$\frac{\partial \xi^*_j}{\partial \alpha} = \sum_{i=1}^{n} \left\{ b_{iy} \frac{\partial P^*_i(\theta_j)}{\partial b^*_{yi}} + \frac{a_{iy} \partial P^*_i(\theta_j)}{\alpha^2 \partial a^*_{yi}} \right\} \qquad [3.5.6]$$

$$\frac{\partial \xi^*_j}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial P^*_i(\theta_j)}{\partial b^*_{yi}} \frac{\partial b^*_{yi}}{\partial \beta} \qquad [3.5.7]$$

where n is the total number of anchor items. The partial derivatives of $P^*_i(\theta_j)$ for the three parameter logistic model are substituted in equations 3.5.6 and 3.5.7. The expanded version of equations 3.5.6 and 3.5.7 are then substituted in equations 3.5.4 and 3.5.5, respectively. Equations 3.5.4 and 3.5.5 are then solved iteratively for α and β in order to minimize the function F (3.5.3).

3.6 Method of Evaluation

A number of different methods have been used to evaluate the results of anchor test scaling studies. Four of the more popular evaluation methods have been outlined by Phillips (1985) and include 1) comparison of scaling results to those obtained from a well established scaling procedure (Lord, 1977; Guskey, 1981), 2) assessment of scale drift (Peterson, Cook & Stocking, 1981), 3) stability of the scalings

using cross-validation groups (Kolen, 1981; Kolen & Whitney, 1982) , and 4) scaling a test to itself (Marco, Petersen & Stewart, 1979; Phillips, 1985).

The first method was not used in this study since this method of evaluation is not appropriate for scaling studies where the effects of several variables on the scaling procedure are to be investigated. Likewise, the second method was not used in this study since multiple editions of a test are not available to allow for evaluating errors in the scaling chain. Methods 1 and 2 are often used for evaluating scaling results when real test data is used. The third method also was not used in this study since evaluating stability does not provide a completely valid criterion for evaluating scaling results. Lord and Wingersky (1984) noted that "Although stability is certainly desirable, stability is not a proper criterion for choosing the best equating method: Incorrect equating procedures may yield more stable results than correct procedures". Evaluating only the stability of scaling results is analogous to evaluating a test through its reliability without consideration of the tests validity. Lastly, the fourth method of evaluating the results of a test scaling study was also not used in this study since scaling a test to itself does not allow for the manipulation of the variables of interest. This method of evaluating scaling results is often used to compare the usefulness of different scaling methods where the characteristics of the test remain constant.

An alternative to these four methods of evaluating scaling results is to compare the scaling results for simulated test data. The advantage of this method is that a number of variables can simultaneously be manipulated. This method of evaluating scaling results has

been used in a number of studies including those by Vale, Maurelli, Gialluca, Weiss, and Ree (1981) and, Wingersky, Cook, and Eignor (1987) and was used to evaluate the results of this study.

More specifically, the results from this simulation study were evaluated in three ways. First, the estimated scaling coefficients were compared to the known true scaling coefficients across various combinations of the two sample sizes, two examinee ability overlaps and four anchor test lengths. These comparisons allowed for a relative evaluation of the effects of these variables on scaling error. Reporting of the scaling coefficients is also useful since the scaling coefficients obtained in other studies are often reported and can be compared to those obtained in this study.

Second, the amount of error from transforming the known and estimated true scores for examinees B on test Y on to the scale for test X using either the true or estimated scaling lines was compared for various combinations of the two sample sizes, two examinee ability overlaps, and four anchor test lengths. Inaccuracy when scaling test scores is the result of item and ability parameter estimation error. This error can produce scaling error in two possible ways. First, there is the effect of parameter estimation error by itself. The location of an examinee's true score on a "base" axis is likely to be different from the location of the examinee's estimated true score on the same axis. Even if the true scaling coefficients were used to scale both sets of true scores, a difference in the scaled scores would result and this difference would reflect the scaling error due to parameter estimation error by itself. This type of scaling error will be referred to as Type I scaling error. Looking at this type of scaling error is useful

because it provides a baseline for interpreting scaling error over and above that which can be expected from simply calibrating the item and ability parameters for a set of data.

Secondly, there is the effect of parameter estimation error on the calculation of the scaling coefficients. With the characteristic curve scaling method, the accuracy of both the item and ability parameter estimates can effect the accuracy of the calculated scaling coefficients. The error associated with scaling an examinee's true score because of error in determining the scaling coefficients, over and above Type I scaling error, will be referred to as Type II scaling error.

These two types of scaling error are shown graphically in Figure 2. The ordinate and abscissa represent the true score scales for tests X and Y. The two lines labeled T and E are the true and estimated scaling lines used to scale the true scores for test Y onto the same scale as test X. Lastly, S and $\hat{S}$ are respectively the "true" and estimated true scores. Because the data used in this study were simulated, the location of the lines T and E and the values of S and $\hat{S}$ were known. The two types of scaling error can be determined by transforming S and $\hat{S}$ onto the scale for test X using the scaling lines T and E and determining the difference between these scaled true scores. The scaled true scores are: (1) S1, S scaled using the true scaling line (T), (2) S2, $\hat{S}$ scaled using the true scaling line (T) and (3) S3, $\hat{S}$ scaled using the estimated scaling line (E). Type I scaling error is the difference between S1 and S2 while Type II scaling error is the difference between S2 and S3.

It can be noted from Figure 2 that two possible cases exist if the intercepts of the scaling lines T and E remain the same. In the first case (shown in Figure 2), $\hat{S}$ is greater than S. In this situation, the

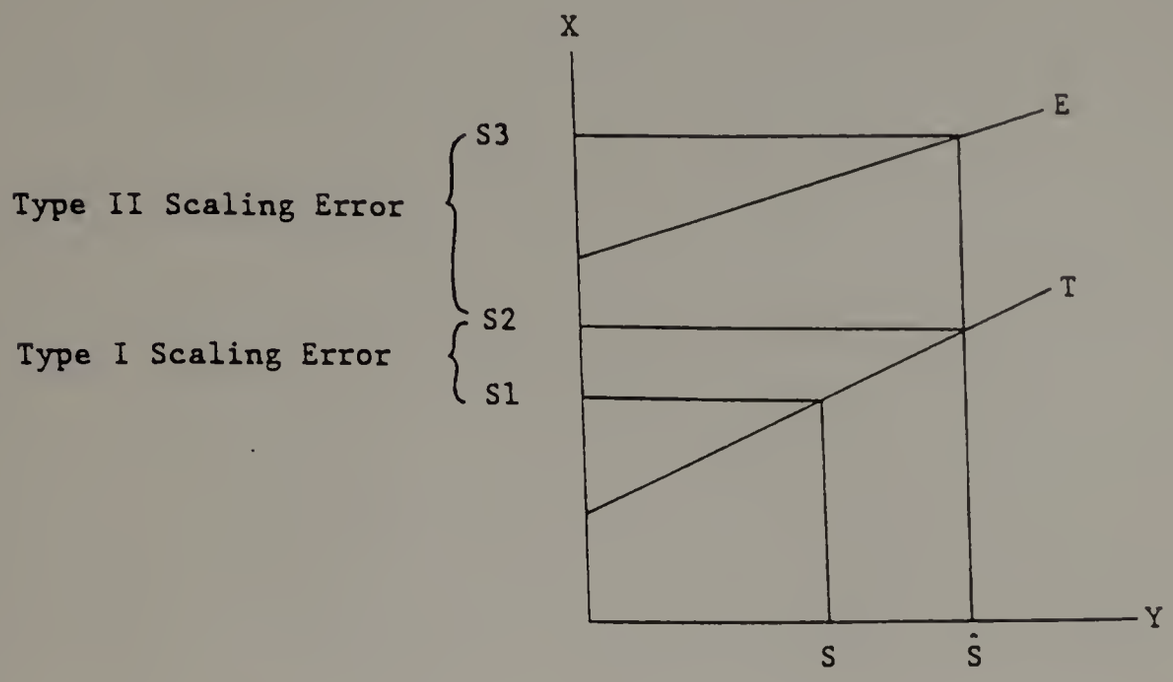

Figure 2. Graphical Representation of Type I and Type II Scaling Error ($\hat{S} > S$)

total scaling error is equal to S3 minus S1 and the Type I scaling error (S2 minus S1) is a subset of the total error. In the second case (shown in figure 3), S is greater than $\hat{S}$ and, in this case, the Type I scaling error (S2 minus S1) is not a subset of the total error. Since it was necessary to compare the results when both case 1 and case 2 were present, the total error used in this study is equal to the sum of the absolute values of the Type I and Type II scaling errors.

The main advantage of evaluating scaling error in this way is that the impact of the results on a test translation study are readily apparent. For example, if, under a specific set of conditions, the mean Type II scaling error is equal to 4.0, this means that, if a test

translation study is conducted and the scaling is performed under similar conditions, then a difference of 4 points can be expected between the mean estimated true score on test X and test Y due to scaling error alone. This difference in estimated true scores can also

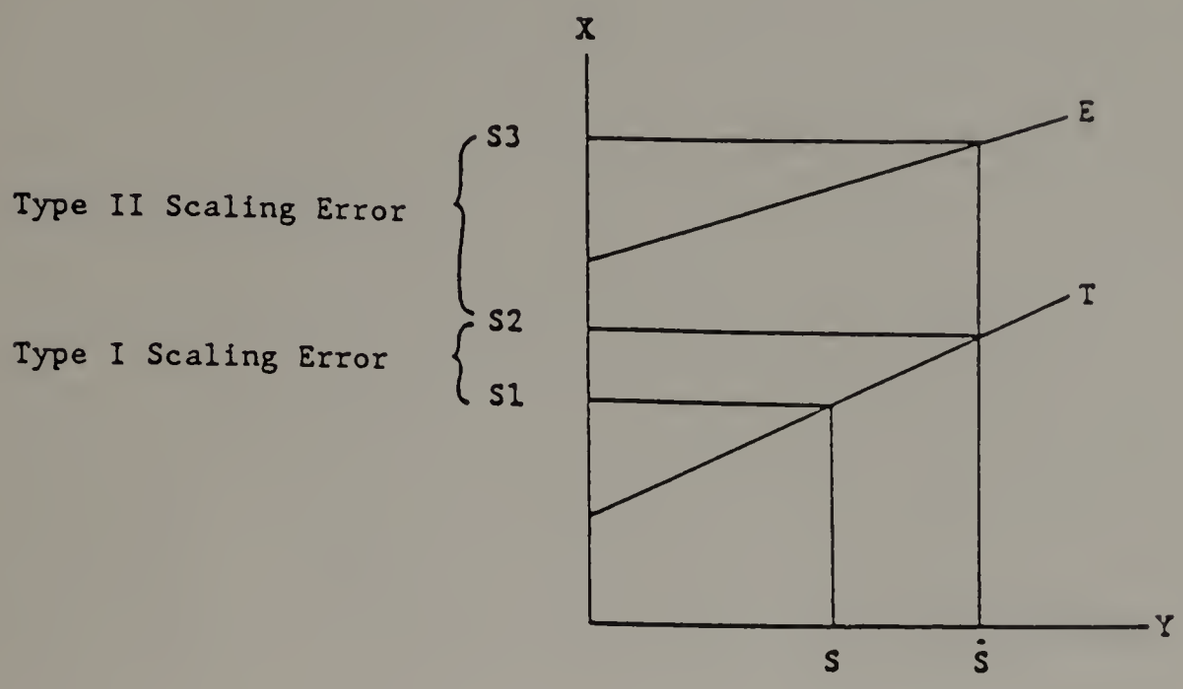

Figure 3. Graphical Representation of Type I and Type II Scaling Error ($\dot{S} < S$)

be thought of in terms of individual items. For a 50-item test, a difference in estimated true scores of 4 points means that, on the average, a difference of .08 in the probability of answering an item on test X and Y can be expected due to scaling error alone. This difference in probabilities may be more for some items and less for others, but, at the least, it provides a rough baseline for interpreting the translation equivalence of test items. This is important since, at this stage of test translation research, understanding the degree of translation equivalence at the item level is important.

In this study, the mean differences in examinee ability (for groups A and B) and item difficulty (for tests X and Y) are 1.35 and 0.51 for the data sets with 50% and 80% ability overlap respectively. Also, the ratio of standard deviations for tests X and Y is 1.0 across all of the data sets. Therefore, the true scaling constants are:

α=1.0 and β=1.35 for the 50% overlap in abilities

α=1.0 and β=0.51 for the 80% overlap in abilities

These true scaling constants were used to obtain the true scaling line (T) shown in Figures 2 and 3.

The third way of evaluating the results from this simulation study was to compare the true and estimated percentile ranks of scaled scores for various combinations of the two sample sizes, two examinee ability overlaps, and four anchor test lengths. These comparisons were carried out by (1) transforming the estimated true scores of examinees A on test X onto the scale for test Y using the true and estimated scaling coefficients and (2) calculating the difference between the percentile ranks obtained from these two sets of transformed estimated true scores. The difference between these percentile ranks reflects the degree of scaling error obtained from inaccurately estimated scaling coefficients. These comparisons are especially useful since they provide a way of evaluating the absolute impact of scaling error in terms of a common way of reporting test scores.

CHAPTER 4

RESULTS

4.1 Introduction

In this chapter, the results of the investigation outlined in Chapter 3 are presented. These results are presented in the context of the evaluation methods discussed in section 3.6, including evaluation of (1) scaling coefficients, (2) Types I and II scaling errors, and (3) change in percentile ranks. The results based on these three evaluation methods are given in sections 4.2, 4.3, and 4.4, respectively. A summary of these results is given in section 4.5.

4.2 Results Based on Scaling Coefficients

The results given in this section are based on the evaluation of scaling coefficients and will be presented in three parts. In the first part, the scaling coefficient results across the two sample sizes and four anchor test lengths will be presented. In the second part, the scaling coefficient results across the two examinee ability overlaps and four anchor test lengths will be presented. Lastly, the scaling results across sample size, ability overlap, and anchor test length will be presented.

4.2.1 Scaling Coefficients Across Sample Size and Anchor Test Length

The estimated scaling coefficients across the two sample sizes and four anchor test lengths are given in Table 5. Also given in Table 5 are the differences and absolute differences between the true and estimated scaling coefficients. These differences in scaling coefficients are referred to as residuals. The true scaling

Table 5

Estimated Scaling Coefficients, Residuals, and Absolute Residuals Across Sample Size and Anchor Test Length (Averaged Across Examinee Ability Overlap)

Sample Size	Anchor Test Length	Scaling Coefficients/Residuals[1]					
		$\hat{\alpha}$	$\alpha-\hat{\alpha}$	$\lvert\alpha-\hat{\alpha}\rvert$	$\hat{\beta}$	$\beta-\hat{\beta}$	$\lvert\beta-\hat{\beta}\rvert$
300	5	1.16	-.16	.30	1.02	-.09	.24
	10	1.17	-.17	.17	.86	.07	.07
	15	1.14	-.14	.14	.88	.05	.05
	20	1.03	-.03	.08	.89	.04	.04
600	5	1.19	-.19	.19	.95	-.02	.08
	10	1.07	-.07	.07	.91	.03	.03
	15	1.06	-.06	.06	.94	-.01	.06
	20	1.02	-.02	.06	.93	.01	.02

[1]The true equating constants are α=1.0 and β=0.93.

coefficients, estimated scaling coefficients, and residuals given in Table 5 were averaged across the results for the 50% and 80% examinee ability overlap samples.

The absolute residual results in Table 5 indicate that there was a greater difference between the true and estimated scaling coefficients for the N=300 samples than for the N=600 samples. This was the case for both the α and β scaling coefficients across the four anchor test lengths with the exception of the β scaling coefficient for the 15-item anchor test sample. This general result of less scaling error with a larger sample size was expected, since doubling an N=300 calibration

sample size should result in substantially more accurate parameter estimation and consequently less scaling error.

As can also be seen from the absolute residual results in Table 5, longer anchor tests generally resulted in less scaling error for both scaling coefficients. This pattern was also expected since, with longer anchor tests, there are more "points" to aid in estimating a scaling line. Therefore, longer anchor tests should result in more accurately estimated scaling coefficients and consequently less scaling error. However, the reduction in scaling error was greatest for the 5- to 10-item increase in anchor test length. Subsequent increases in anchor test length had a relatively minor effect on the reduction of scaling error.

4.2.2 Scaling Coefficients Across Examinee Ability Overlap and Anchor Test Length

The estimated scaling coefficients and residuals across the two levels of examinee ability overlap and four anchor test lengths are given in Table 6. The scaling coefficients and residuals in Table 6 were averaged across the results for the N=300 and N=600 samples.

The absolute residual results for α and β in Table 6 indicate that in general there was a greater difference between the true and estimated scaling coefficients for the 50% examinee ability overlap samples than for the 80% ability overlap samples. The few exceptions to this general pattern were the β coefficient residuals for the 5-, 10-, and 15-item anchor test samples. In these cases, the residuals for the 50% ability overlap samples were particularly small compared to those for the 80% ability overlap samples. These low β coefficient residuals will be discussed in conjunction with Table 7. This general pattern

Table 6

Estimated Scaling Coefficients, Residuals, and Absolute Residuals Across Examinee Ability Overlap and Anchor Test Length (Averaged Across Sample Size)

Examinee Ability Overlap	Anchor Test Length	Scaling Coefficients/Residuals[1]					
		$\hat{\alpha}$	$\alpha-\hat{\alpha}$	$\lvert\alpha-\hat{\alpha}\rvert$	$\hat{\beta}$	$\beta-\hat{\beta}$	$\lvert\beta-\hat{\beta}\rvert$
50%	5	1.37	-.37	.37	1.56	-.21	.21
	10	1.19	-.19	.19	1.34	.01	.01
	15	1.15	-.15	.15	1.37	-.02	.04
	20	1.09	-.09	.09	1.36	-.01	.01
80%	5	0.98	.02	.12	.41	.11	.11
	10	1.05	-.05	.05	.43	.09	.09
	15	1.05	-.05	.05	.45	.07	.07
	20	0.96	.05	.05	.46	.05	.05

[1]The true equating constants are α=1.0, β=1.35 for the 50% examinee ability overlap and α=1.0, β=0.51 for the 80% examinee ability overlap.

of less scaling error with a greater examinee ability overlap was expected. With a greater examinee ability overlap, there were a greater number of examinees located in the region of ability where the anchor items were located and therefore the parameters for these items are estimated more accurately than they would be with a smaller examinee ability overlap. More accurately estimated item parameters should result in less scaling error.

The absolute residual results given in Table 6 also indicate that an increase in anchor test length from 5 to 10 items resulted in reduced scaling error. This reduction in scaling error was particularly

Table 7

Estimated Scaling Coefficients, Residuals, and Absolute Residuals Across Sample Size, Examinee Ability Overlap, and Anchor Test Length

Sample Size	Examinee Ability Overlap	Anchor Test Length	Scaling Coefficients/Residuals[1]					
			$\hat{\alpha}$	$\alpha-\hat{\alpha}$	$\lvert\alpha-\hat{\alpha}\rvert$	$\hat{\beta}$	$\beta-\hat{\beta}$	$\lvert\beta-\hat{\beta}\rvert$
300	50%	5	1.45	-.45	.45	1.68	-.33	.33
		10	1.25	-.25	.25	1.34	.01	.01
		15	1.19	-.19	.19	1.33	.02	.02
		20	1.10	-.10	.10	1.35	.00	.00
	80%	5	.86	.14	.14	.36	.15	.15
		10	1.09	-.09	.09	.38	.13	.13
		15	1.08	-.08	.08	.43	.08	.08
		20	.95	.05	.05	.43	.08	.08
600	50%	5	1.28	-.28	.28	1.44	-.09	.09
		10	1.13	-.13	.13	1.34	.01	.01
		15	1.10	-.10	.10	1.41	-.06	.06
		20	1.08	-.08	.08	1.36	-.01	.01
	80%	5	1.10	-.10	.10	.45	.06	.06
		10	1.00	.00	.00	.47	.04	.04
		15	1.02	-.02	.02	.46	.05	.05
		20	.96	.04	.04	.49	.02	.02

[1]The true equating constants are α=1.0, β=1.35 for the 50% examinee ability overlap and α=1.0, β=0.51 for the 80% examinee ability overlap.

evident for the 50% examinee ability overlap samples. Subsequent increases in anchor test length reduced the scaling error for the α

scaling coefficient with the 50% ability overlap samples and for the β scaling coefficient with the 80% ability overlap samples. However, these reductions in scaling error were relatively minor compared to those for the 5- to 10-item increase in anchor test length.

4.2.3 Scaling Coefficients Across Sample Size, Examinee Ability Overlap and Anchor Test Length

The scaling coefficients and residuals across sample size, examinee ability overlap, and anchor test length are given in Table 7 (p. 100). These absolute residual results follow the same general trends discussed previously. These general trends for both scaling coefficients include (1) greater scaling error for the N=300 samples than for the N=600 samples, (2) greater scaling error for the 50% examinee ability overlap samples than for the 80% ability overlap samples, and (3) the greatest reduction in scaling error with an increase in anchor test length of 5 to 10 items with relatively small reductions in scaling error for subsequent increases in anchor test length.

Table 7 also provides further insights into these general trends. First, the reduction in scaling error for the α scaling coefficient with increased examinee ability overlap was greater for the N=300 samples than for the N=600 samples. This was not the case for the β scaling coefficient. However, the results given in Table 7 indicate that the accuracy of the β scaling coefficients for the 5-, 10-, and 15-item anchor test samples within the 50% examinee ability overlap for the N=300 sample sizes should be questioned. Compared to the residuals for the corresponding samples within the 80% examinee ability overlap, the β coefficient residuals for these samples were extremely low. Even though

these results were based on five replications, they were still subject to sampling error that could have produced inconsistent results. The scaling results for the β coefficients for the N=600 samples were quite good overall, suggesting a ceiling effect for this scaling coefficient with these samples. It can be seen from Table 7 that the β coefficients for the eight N=600 samples showed little variance and that they were estimated more accurately for the N=600 than for the N=300 samples. This indicates that varying the anchor test length or examinee ability overlap had a minimal impact in the resulting β coefficients for the N=600 samples. Second, the reduction in scaling error for the α scaling coefficient with an increase in anchor test length from 5 to 10 items was greatest for the N=300 and 50% examinee ability overlap samples and least for the N=300 and 80% examinee ability overlap samples. These results for the β scaling coefficient are less interpretable because of the reasons stated above. Lastly, for the N=300 and 50% examinee ability overlap samples, relatively large reductions in scaling error for the α scaling coefficients were obtained with subsequent increases in anchor test length compared to the other samples. Again, these results for the β scaling coefficient are less interpretable because of the reasons stated previously.

In summary, these results indicate that across all of the samples, 5-item anchor tests result in the most scaling error. Increasing the anchor test length to 10 items substantially reduced the scaling error across the four samples, but was particularly helpful for the 50% ability overlap sample. Also, with the exception of the N=300 and 50% examinee ability overlap samples, increasing the anchor test length beyond 10 items had a relatively minor impact on the reduction of

scaling error. In general, larger sample sizes and greater examinee ability overlaps resulted in less scaling error across the four anchor test lengths.

4.3 Results Based on Type I and Type II Scaling Error

As was the case in section 4.2, the results in this section will be presented in three parts. In the first part, the Type I, Type II, and total scaling errors across the two sample sizes and four anchor test lengths will be presented. In the second part, these scaling errors across the two ability overlaps and four anchor test lengths will be presented. Lastly, these scaling errors across sample size, ability overlap, and anchor test length will be presented.

As discussed in Section 3.6, Type I scaling error is the error associated with scaling estimated true scores using the known true scaling line. It reflects the amount of scaling error that can be expected from parameter estimation error alone. Type II scaling error is the error associated with scaling estimated true scores using the estimated scaling line. It reflects the amount of scaling error that can be expected from the effects of parameter estimation error on the calculation of the scaling coefficients.

The Type I, Type II, and total scaling error results in the following three sections are given in terms of mean difference (MD), mean absolute difference (MAD), and root mean squared difference (RMSD). The mean absolute difference is particularly relevant for interpreting these results since it reflects the amount of scaling error that could be obtained if the error were unidirectional. Mean absolute difference, therefore, represents the maximum amount of scaling error that could have been obtained. Root mean squared difference is equal to the square

root of the sum of the squared scaling error. This index adds weight over the mean absolute difference when more extreme differences are present. The root mean squared difference is often reported in the scaling literature either as given here or as a weighted difference that takes into account the frequencies of the examinees' scores.

4.3.1 Type I, Type II, and Total Scaling Error Across Sample Size and Anchor Test Length

The Type I scaling error across the two sample sizes and four anchor test lengths is given in Table 8. These results indicate that the N-300 samples had greater MAD scaling error than the N-600 samples. These results also indicate that increasing the anchor test length reduced the MAD scaling error for the N-300 samples, but not for the N-600 samples. For the N-600 samples, a .05 decrease in MAD scaling error was obtained when the anchor test was increased from 5 to 10 items. However, subsequent increases in anchor test length failed to decrease the MAD scaling error below the 10 anchor item level.

One possible explanation for this result is a ceiling effect. It can be seen from Table 8 that the MAD scaling error for the four N-600 samples showed little variance and there was less MAD scaling error for the N-600 samples than for the N-300 samples. This indicates that the results for the N-600 samples were quite good and that varying the anchor test length made little discernable difference in the resulting Type I scaling error.

It should also be noted that, for the N-300 samples, the reduction in scaling error was greatest for an anchor test length increase of 5 to 10 items and least for an increase of 10 to 15 items. The increase in

Table 8

Type I Scaling Error Across Sample Size and Anchor Test Length (Averaged Across Examinee Ability Overlap)

Sample Size	Anchor Test Length	Type I Scaling Error[1]		
		MD	MAD	RMSD
300	5	.37	.80	.87
	10	.42	.49	.64
	15	.05	.46	.56
	20	.08	.32	.37
600	5	.06	.28	.34
	10	.03	.23	.27
	15	-.04	.29	.35
	20	-.03	.26	.42

[1] MD = Mean Difference
MAD = Mean Absolute Difference
RMSD = Root Mean Squared Difference

anchor test length from 15 to 20 items resulted in a medium reduction in scaling error.

The corresponding results for Type II scaling error are given in Table 9. As was the case with the Type I scaling error, the N=300 samples had greater MAD scaling error than the N=600 samples. This difference in scaling error between the two sample sizes was greater than the differences for Type I scaling error. Three additional trends are also evident from Table 9. First, for both sample sizes, increasing the anchor test length reduced the scaling error. Second, the greatest reduction in scaling error with increased anchor test length occurred

Table 9

Type II Scaling Error Across Sample Size and Anchor Test Length (Averaged Across Examinee Ability Overlap)

Sample Size	Anchor Test Length	Type II Scaling Error[1]		
		MD	MAD	RMSD
300	5	.47	2.20	2.81
	10	-.24	1.04	1.19
	15	-.31	.83	1.03
	20	-.18	.58	.64
600	5	.34	1.07	1.28
	10	-.01	.53	.63
	15	.08	.50	.62
	20	.02	.36	.42

[1] MD - Mean Difference
MAD - Mean Absolute Difference
RMSD - Root Mean Squared Difference

with the N=300 samples. Third, the greatest reduction in scaling error for both sample sizes occurred when the anchor test was increased from 5 to 10 items and the least reduction occurred with an increase from 10 to 15 items. It should also be noted that, for each sample, the Type II scaling error was greater than the Type I scaling error.

Table 10 gives the total scaling error across the two sample sizes and four anchor test lengths. The total scaling error is equal to the sum of the Type I and Type II scaling error and therefore represents the maximum amount of scaling error that can be expected under conditions similar to those used in this study.

Table 10

Total Scaling Error Across Sample Size and Anchor Test Length (Averaged Across Examinee Ability Overlap)

Sample Size	Anchor Test Length	Total Scaling Error[1] MAD	RMSD
300	5	3.00	3.68
	10	1.53	1.83
	15	1.29	1.59
	20	.90	1.01
600	5	1.35	1.62
	10	.76	.90
	15	.79	.97
	20	.62	.84

[1] MAD = Mean Absolute Difference
RMSD = Root Mean Squared Difference

Since the total scaling error is a function of both Type I and Type II scaling error, the general trends exhibited in the previous two tables are evident here as well. These trends include (1) greater scaling error for the N=300 samples than for the N=600 samples, (2) less scaling error with increased anchor test length for both sample sizes, (3) greater reduction in scaling error with increased anchor test length for the N=300 samples than for the N=600 samples, and (4) the greatest reduction in scaling error with an anchor test length increase from 5 to 10 items and the least reduction with an increase of from 10 to 15 items.

The results given in Table 10 can also be interpreted in absolute terms. An increase in the sample size from 300 to 600 reduced the MAD scaling error from 3.00 to 1.35 for the 5-item anchor test. A scaling error of 3.00 on a 50-item test represents a 6.00% scaling error. This is a fairly substantial scaling error. A reduction in the scaling error to 2.70% by an increase in the sample size from 300 to 600 is certainly a helpful reduction in scaling error. The reduction in scaling error with the same increase in sample size was less for longer anchor test lengths, with the least reduction obtained with the 20-item anchor test. With the 20-item anchor test, the scaling error was reduced from 1.80% for the N=300 sample to 1.24% for the N=600 sample.

4.3.2 Type I, Type II, and Total Scaling Error Across Examinee Ability Overlap and Anchor Test Length

The Type I scaling error across the two levels of examinee ability overlap and four anchor test lengths are given in Table 11. These results indicate that the 50% ability overlap samples had greater MAD scaling error than the 80% ability overlap samples. These results also indicate that increasing the anchor test length generally reduced the scaling error for both the 50% and 80% ability overlap samples. The exception for all of the samples in both ability overlaps was with an anchor test increase from 10 to 15 items. In these cases, the MAD increased slightly. It can also be seen from Table 11 that the general reduction in scaling error with increased anchor test length was greater for the 50% examinee ability overlap samples than for the 80% ability overlap samples. Lastly, for the 50% ability overlap samples, the greatest reduction in scaling error occurred with an increase in anchor test length from 5 to 10 items.

Table 11

Type I Scaling Error Across Examinee Ability Overlap and Anchor Test Length (Averaged Across Sample Size)

Examinee Ability Overlap	Anchor Test Length	Type I Scaling Error[1]		
		MD	MAD	RMSD
50%	5	.09	.64	.71
	10	.22	.33	.39
	15	-.13	.35	.45
	20	.08	.26	.31
80%	5	.34	.44	.51
	10	.23	.39	.52
	15	.13	.40	.46
	20	-.03	.32	.39

[1] MD - Mean Difference
MAD - Mean Absolute Difference
RMSD - Root Mean Squared Difference

The corresponding results for Type II scaling error are given in Table 12. These results are similar to those obtained for Type I MAD scaling error, except (1) the scaling error is larger in all cases, (2) there is a consistent trend of reduced scaling error with increased anchor test length and (3) the greatest reduction in scaling error occurred with an increase in anchor test length from 5 to 10 items for the samples in both ability overlaps.

Table 13 gives the total scaling error across examinee ability overlap and anchor test length. These results exhibit the same general trends as the previous two tables and include (1) greater scaling error

Table 12

Type II Scaling Error Across Examinee Ability Overlap and Anchor Test Length (Averaged Across Sample Size)

Examinee Ability Overlap	Anchor Test Length	Type II Scaling Error[1]		
		MD	MAD	RMSD
50%	5	1.54	2.33	2.89
	10	.25	1.04	1.18
	15	.33	.87	1.02
	20	.09	.53	.59
80%	5	-.74	.94	1.20
	10	-.49	.53	.64
	15	-.41	.45	.63
	20	-.26	.40	.47

[1] MD - Mean Difference
MAD - Mean Absolute Difference
RMSD - Root Mean Squared Difference

for the 50% ability overlap samples than for the 80% ability overlap samples, (2) less scaling error with increased anchor test length for the samples in both ability overlaps, (3) greater reduction in scaling error with increased anchor test length for the 50% ability overlap samples than for the 80% ability overlap samples, and (4) the greatest reduction in scaling error with an anchor test length increase from 5 to 10 items and the least reduction with an increase from 10 to 15 items.

Again, it is interesting to interpret these scaling error results in absolute terms. For the results in Table 13, an increase in the examinee ability overlap from 50% to 80% reduced the MAD scaling error

Table 13

Total Scaling Error Across Examinee Ability Overlap and Anchor Test Length (Averaged Across Sample Size)

Examinee Ability Overlap	Anchor Test Length	Total Scaling Error[1] MAD	RMSD
50%	5	2.97	3.60
	10	1.37	1.57
	15	1.22	1.47
	20	.79	0.90
80%	5	1.38	1.71
	10	.92	1.16
	15	.85	1.09
	20	.72	.86

[1] MAD = Mean Absolute Difference
RMSD = Root Mean Squared Difference

from 2.97 to 1.38 for the 5-item anchor test. For a 50-item test, these scaling errors represent a 5.94% and 2.76% scaling error respectively. The reduction in scaling error with the same increase in ability overlap was less for longer anchor test lengths, yet the reduction was still substantial with the exception of the 20-item anchor test. With the 20-item anchor test, the scaling error was 1.58% for the 50% ability overlap sample and 1.44% for the 80% ability overlap sample.

4.3.3 Type I, Type II, and Total Scaling Error Across Sample Size, Examinee Ability Overlap, and Anchor Test Length

The Type I, Type II, and total scaling error across sample size, examinee ability overlap and anchor test length are given in Tables 14

through 16, respectively. The results given in each of these tables follow the same trends and therefore they will be discussed simultaneously.

The results given in Tables 14 through 16 exhibit the same general trends that have been discussed previously. These general trends include (1) greater scaling error for the N=300 sample than for the N=600 sample, (2) greater scaling error for the 50% ability overlap sample than for the 80% ability overlap sample, (3) less scaling error with increased anchor test length for both sample sizes and examinee ability overlaps, (4) greater reduction in scaling error with an increased anchor test length for the N=300 sample than for the N=600 sample and for the 50% ability overlap than for the 80% ability overlap, and (5) the greatest reduction in scaling error with an anchor test length increase from 5 to 10 items and the least reduction with an increase from 10 to 15 items for both sample sizes and ability overlaps.

The MAD scaling results given in Tables 14 through 16 provide further insights into these general trends. First, the reduction in total scaling error with an increase in examinee ability overlap was generally greater for the N=300 samples than for the N=600 samples. This was the case across all of the anchor test lengths with the exception of the 20-item anchor test for the combined N=300 and 50% examinee ability overlap sample. Second, the reduction in scaling error with an increase in sample size for each ability overlap was greater than the reduction in scaling error with an increase in ability overlap for each sample size (this was particularly the case for the Type I scaling error). This result indicates that, for the sample sizes and examinee ability overlaps used in this study, sample size plays a

Table 14

Type I Scaling Error Across Sample Size, Examinee Ability Overlap, and Anchor Test Length

Sample Size	Examinee Ability Overlap	Anchor Test Length	Type I Scaling Error [1]		
			MD	MAD	RMSD
300	50%	5	.16	1.00	1.08
		10	.35	.46	.58
		15	- .10	.44	.57
		20	.13	.18	.22
	80%	5	.57	.59	.66
		10	.49	.52	.70
		15	.19	.48	.54
		20	.03	.45	.52
600	50%	5	.02	.27	.33
		10	.08	.19	.20
		15	- .15	.25	.33
		20	.02	.33	.39
	80%	5	.10	.29	.35
		10	- .03	.26	.33
		15	.07	.32	.37
		20	- .08	.19	.25

[1]MD - Mean Difference
MAD - Mean Absolute Difference
RMSD - Root Mean Squared Difference

Table 15

Type II Scaling Error Across Sample Size, Examinee Ability Overlap, and Anchor Test Length

Sample Size	Examinee Ability Overlap	Anchor Test Length	Type II Scaling Error [1]		
			MD	MAD	RMSD
300	50%	5	2.07	2.89	3.64
		10	.23	1.30	1.43
		15	.21	1.05	1.16
		20	.03	.62	.67
	80%	5	-1.14	1.50	1.98
		10	- .70	.77	.95
		15	- .52	.60	.90
		20	- .38	.53	.62
600	50%	5	1.00	1.76	2.14
		10	.27	.77	.92
		15	.45	.69	.87
		20	.14	.44	.51
	80%	5	- .33	.37	.41
		10	- .28	.28	.33
		15	- .30	.30	.36
		20	- .13	.27	.33

[1]MD - Mean Difference
MAD - Mean Absolute Difference
RMSD - Root Mean Squared Difference

Table 16

Total Scaling Error Across Sample Size, Examinee Ability Overlap, and Anchor Test Length

Sample Size	Examinee Ability Overlap	Anchor Test Length	Total Scaling Error[1] MAD	Total Scaling Error[1] RMSD
300	50%	5	3.89	4.72
		10	1.76	2.01
		15	1.49	1.73
		20	.80	.89
	80%	5	2.09	2.64
		10	1.29	1.65
		15	1.08	1.44
		20	.98	1.14
600	50%	5	2.03	2.47
		10	.96	1.12
		15	.94	1.20
		20	.77	.90
	80%	5	.66	.76
		10	.54	.66
		15	.62	.73
		20	.46	.58

[1]MAD = Mean Absolute Difference
RMSD = Root Mean Squared Difference

greater role than ability overlap in the amount of scaling error that is obtained. Third, the reduction in scaling error with an increase in anchor test length from 5 to 10 items was greatest for the N-300 and 50% examinee ability overlap sample and least for the N-600 and 80% examinee ability overlap sample. The N-300 and 80% ability overlap sample and the N-600 and 50% ability overlap sample showed a substantial reduction in scaling error for this increase in anchor test length.

Fourth, increases in anchor test length from 15 to 20 items had a relatively minor overall impact on reducing scaling error. This was especially true for the N-600 samples. In fact, for the N-600 and 80% examinee ability overlap sample, the scaling error actually increased slightly with an increase in anchor test length from 15 to 20 items. Increasing the anchor test length from 15 to 20 items had a minimal impact on reducing scaling error for all of the samples with the exception of the N-300 and 50% examinee ability overlap sample. With this sample, this increase in anchor test length reduced the total MAD scaling error by approximately one half. However, as was mentioned previously, the Type I scaling error obtained for this sample with a 20-item anchor test was abnormally low, resulting in an abnormally low total scaling error as well.

Fifth, the scaling error for the N-600 and 80% examinee ability overlap sample was quite low across the four anchor test lengths. Even with the 5-item anchor test, the total MAD scaling error was 0.66, which corresponds to a 1.32% scaling error for a 50-item test. When 20 anchor items were used with this sample, the total MAD scaling error was 0.46, which corresponds to an 0.92% scaling error. These results suggest that, in an ideal situation where a "large" sample size is used and the

examinee ability overlap is also large, shorter anchor tests provide nearly comparable scaling results to those obtained with larger anchor tests.

Lastly, it can be seen from Tables 14 and 15 that the majority of the total scaling error is due to Type II rather than Type I scaling error.

In summary, the results given in section 4.3 indicate that larger sample sizes and greater examinee ability overlap resulted in less scaling error. The general reduction in scaling error with an increase in examinee ability overlap was greater for the N=300 samples than for the N=600 samples. Also, across all of the samples, 5-item anchor tests resulted in the most scaling error. Increasing the anchor test length to 10 items substantially reduced the scaling error for the four samples with the exception of the N=600 and 80% examinee ability overlap sample, and was especially helpful for the N=300 and 50% examinee ability overlap sample. Increasing the anchor test length beyond 10 items had a relatively minor impact on the reduction of scaling error.

Furthermore, these results indicate that, for the sample sizes and examinee ability overlaps used in this study, sample size plays a greater role than ability overlap in the amount of scaling error that is obtained through item calibration alone. In addition, under ideal scaling conditions of larger sample sizes and a large examinee ability overlap (such as the N=600 and 80% examinee ability overlap samples used in this study), shorter anchor tests provide nearly comparable scaling results to those obtained with longer anchor tests. Lastly, the majority of the scaling error obtained was due to Type II scaling error

(scaling error over and above parameter estimation error) rather than Type I scaling error.

4.4 Results Based on Change in Percentile Ranks

The third way of evaluating the results from this study was through change in percentile ranks. Test scores are often used to rank examinees relative to each other on a trait of interest and percentile ranks are a common way of reporting these scores. For example, results from the College Board Scholastic Aptitude Test, which has been translated for use with Spanish-speaking populations, are often reported in terms of percentile ranks. For this reason, it was of interest to investigate the effects of scaling error due to sample size, examinee ability overlap, and anchor test length on examinees' percentile ranks.

The results in this section will be presented in two parts. In the first part, the change in percentile ranks across the two sample sizes and the four anchor test lengths will be presented. In the second part, the change in percentile ranks across sample size, examinee ability overlap, and anchor test length will be presented.

4.4.1 Change in Percentile Ranks Across Sample Size and Anchor Test Length

The absolute residuals of percentile ranks across sample size and anchor test length for various test scores are given in Table 17. These residuals were calculated by subtracting the percentile ranks obtained using estimated equating constants from the percentile ranks obtained using the true equating constants. Absolute values of these residuals were then averaged across the 50% and 80% examinee ability overlap samples. Absolute values of the percentile ranks were used to consider the maximum amount of scaling error that could have been

Table 17

Absolute Residuals of Percentile Ranks Across Sample Size and Anchor Test Length for Various Test Scores (Averaged Across Examinee Ability Overlap)

Sample Size	Anchor Test Length	Absolute Residuals of Percentile Ranks for Various Test Scores						
		15	20	25	30	35	40	45
300	5	0.7	2.0	8.5	12.9	12.7	9.5	4.7
	10	1.8	4.2	5.7	4.0	5.2	2.5	1.2
	15	1.7	4.0	4.2	3.2	3.0	1.8	0.5
	20	0.2	1.5	1.4	2.3	3.5	4.0	0.9
600	5	1.6	2.0	4.3	7.6	5.7	4.5	1.9
	10	1.4	1.8	1.0	3.4	2.8	2.1	0.6
	15	0.7	1.5	3.2	4.9	3.3	2.7	0.5
	20	0.4	1.1	0.1	1.3	1.7	2.2	1.1

obtained, and to allow for the meaningful averaging of both positive and negative values across the examinee ability overlaps in each sample. The residuals of the percentile ranks for scores of 0, 5, 10, and 50 were 0.0 across all 16 samples. Residuals for these scores are not listed in Table 17 and likewise are not listed in Table 18.

The absolute residuals given in Table 17 vary between the different test scores. They are generally higher for the test scores of 30 and 35 than for the remaining test scores for each of the samples. These higher absolute residuals for test scores of 30 and 35 are the result of having a greater number of examinees located in this region of the true score scale. Slight changes in an examinee's true score due to scaling error will have a more profound effect on an examinee's percentile rank in this score region compared to other score regions where the number of examinees is smaller.

Table 18

Residuals of Percentile Ranks Across Sample Size, Examinee Ability Overlap, and Anchor Test Length for Various Test Scores

Sample Size	Examinee Ability Overlap	Anchor Test Length	Residuals of Percentile Ranks for Various Test Scores						
			15	20	25	30	35	40	45
300	50%	5	1.3	1.7	-10.0	-15.7	-16.7	-12.7	-5.3
		10	2.3	3.3	4.3	0.0	-4.0	-2.7	-2.3
		15	2.7	5.7	1.7	-1.7	-2.3	-2.3	-1.0
		20	0.3	1.3	1.0	-0.3	-0.3	-3.3	-1.0
	80%	5	0.0	2.3	7.0	10.0	8.7	6.3	4.0
		10	1.3	5.0	7.0	8.0	6.3	2.3	0.0
		15	0.7	2.3	6.7	4.7	3.7	1.3	0.0
		20	0.0	1.7	2.7	4.3	6.7	4.7	0.7
600	50%	5	2.5	2.7	-3.5	-8.9	-8.0	-5.8	-2.7
		10	2.2	3.0	0.0	-2.5	-3.0	-1.9	-1.0
		15	0.7	1.4	-1.2	-6.2	-5.0	-3.8	-1.0
		20	0.5	2.0	0.0	-1.4	-1.2	-2.2	-1.2
	80%	5	0.7	1.2	5.0	6.2	3.3	3.1	0.1
		10	0.5	0.5	2.0	4.2	2.5	2.2	0.2
		15	0.6	1.6	5.2	3.5	1.5	1.5	0.0
		20	-0.2	-0.2	0.2	1.2	2.1	2.1	1.0

The results given in Table 17 reflect the general trends obtained in sections 4.2 and 4.3. First, there was greater scaling error for the N=300 samples than for the N=600 samples. Second, increasing the anchor test length from 5 to 10 items substantially reduced the scaling error while subsequent increases in anchor test length had less impact on reducing scaling error. This was the case across the four anchor test lengths with the exception of the 15-item anchor test for the N=600 sample. For the N=600 sample, the absolute residuals across a number of the test scores for the 15-item anchor test actually increased compared

to those for the 10-item anchor test. This inconsistent result is a function of inconsistencies in calculating the scaling coefficients α and β as mentioned in section 4.2.

A problem with using percentile ranks to report test scores is that, when the variability of the scores is low, small differences in the scores of examinees can lead to large differences in the examinees' percentile ranks. In the context of this study, the same differences in examinees' scores would lead to different differences in the examinee's percentile ranks, depending on whether a 50% or 80% examinee ability distribution were used. Because of this, scaling results across the two ability distributions were not directly compared.

4.4.2 Change in Percentile Ranks Across Sample Size, Examinee Ability Overlap, and Anchor Test Length

The residuals of percentile ranks across sample size, examinee ability overlap, and anchor test length for various test scores are given in Table 18. These residuals are signed and, as can be seen from this Table, both positive and negative residuals were obtained. However, the absolute values of these residuals are of particular interest since they represent the maximum amount of scaling error that could have been obtained if the residuals were unidirectional. The results given in Table 18 are broken down by examinee ability overlap, but, as mentioned previously, comparisons of results across examinee ability overlaps will not be presented.

In addition to exhibiting the same general trends as Table 17, the results given in Table 18 indicate that, for both the 50% and 80% examinee ability overlap samples, the reduction in absolute residual scaling errors with an increase in anchor test length from 5 to 10 items

was greater for the N=300 samples than for the N=600 samples. The results concerning subsequent increases in anchor test length were varied. For the N=300 and 50% examinee ability overlap sample, the increase in anchor test length from 10 to 15 items resulted in small overall reductions in absolute residual scaling error, while the increase in anchor test length from 15 to 20 items resulted in more substantial overall reductions in absolute residual scaling error. This final reduction in scaling errors is helpful but relatively small compared to the reductions obtained with the initial 5-item increase in anchor test length. For the N=300 and 80% examinee ability overlap sample, the increase in anchor test length from 10 to 15 items substantially reduced the absolute residual scaling errors. The final increase in anchor test length to 20 items resulted in substantial increases or decreases in the absolute residual scaling errors for different test scores.

For the N=600 samples, there was an increase in the overall absolute residual scaling error with an increase in anchor test length from 10 to 15 items, as described in reference to Table 17. Increasing the anchor test length from 15 to 20 items substantially reduced the overall residual scaling error for the N=600 samples. However, this overall reduction in scaling error is small when the spurious results obtained for the N=600 samples with the 15-item anchor test are considered. This reduction in scaling errors was greater for the N=600 and 50% examinee ability overlap samples than for the N=600 and 80% examinee ability overlap samples.

Perhaps the more interesting aspect of Table 18 is that it provides for evaluation of scaling error in absolute terms. It can be

seen from Table 18 that the use of a 5-item anchor test with an N=300 sample that has a 50% examinee ability overlap can result in substantial error in examinees' percentile ranks due to scaling error. With this sample and under the conditions used in this study, percentile ranks of examinees with a score of 35 on a translated test can be off by 16.7 percentage points due to scaling error alone. This degree of error is very substantial. It is especially substantial when conducting a test translation study since there are many potential sources of error in addition to scaling error that may occur. The residual percentile ranks given in Table 18 for a number of other test scores and samples are also quite substantial. For the 50% and 80% examinee ability overlap samples, scaling error can generally be reduced to more acceptable levels by assuring that minimum anchor test lengths of approximately 10 items are used.

In summary, the results given in section 4.4 indicate that larger sample sizes generally resulted in less scaling error. These results also indicate that, across all of the samples, 5-item anchor tests resulted in the most scaling error. Increasing the anchor test length to 10 items substantially reduced the scaling error for the four samples with the exception of the N=600 and 80% examinee ability overlap sample, and was particularly helpful for the N=300 and 50% examinee ability overlap sample. Increasing the anchor test length to 15 items resulted in overall reductions in scaling error for the two N=300 samples and some overall increase in scaling error for the two N=600 samples. Increasing the anchor test length to 20 items resulted in relatively small reductions in scaling error overall, particularly for the 80% examinee ability overlap samples. Lastly, the amount of scaling error

obtained varied with the test score with greater scaling error for scores in the 30 to 35 region.

4.5 Summary of Results

The following is a summary of the scaling error results obtained using one or more of the three evaluation methods. Each result will be followed by an (a) if the result was obtained through the evaluation of scaling coefficients, a (b) if the result was obtained through the evaluation of Type I and Type II scaling error, or a (c) if the result was obtained through the evaluation of change in percentile rank. Exceptions to these results may have been obtained because of sampling error, but for the evaluations methods noted, these results were generally obtained. These results are:

1. Greater scaling error for the N=300 samples than for the N=600 samples. (a),(b),(c)
2. Greater scaling error for the 50% examinee ability overlap samples than for the 80% examinee ability overlap samples. (a),(b)
3. Similar reductions in scaling error with an increase in sample size for the 50% examinee ability overlap samples as for the 80% examinee ability overlap samples. (a),(b)
4. Greater reductions in scaling error with an increase in examinee ability overlap for the N=300 samples than for the N=600 samples. (a),(b)
5. The reduction in scaling error with an increase in sample size for a given ability overlap was greater than the reduction in scaling error obtained with an increase in ability overlap for a given sample size. (b) (c)

6. The reduction in scaling error with an increase in anchor test length from 5 to 10 items was large for the N-300 and 50% examinee ability overlap samples and small for the N-600 and 80% ability overlap samples. The N-300 and 80% ability overlap samples and the N-600 and 50% ability overlap samples showed substantial reduction in scaling error with this 5 item increase in anchor test length. (a),(b),(c)

7. Overall, increases in anchor test length from 10 to 15 items had a minimal impact on reducing the scaling error for all of the samples. (a),(b),(c)

8. Increasing the anchor test length from 15 to 20 items generally had a small impact on reducing the scaling error for all of the samples. (a),(b),(c)

These eight general results were obtained using all three of the evaluation methods with only a few exceptions. Results 2, 3 and 4 were not evaluated using method c. Also, result 5 was only obtained with methods b and c. In addition, of the total scaling error obtained through the evaluation of Types I and II scaling error (method b), the majority was due to Type II scaling error, though the amount of Type I error was substantial.

Conclusions based on these results are presented in Chapter 5.

CHAPTER 5

CONCLUSIONS

One focus of this thesis was to provide a review of the history, problems and techniques associated with establishing the translation equivalence of tests. As in previous discussions, the term tests also refers to questionnaires and inventories. Tests have been and will likely continue to be translated into languages that are more suitable for target populations. Historically, the incentive for translating tests was either economic pressure or lack of available testing expertise to develop tests in a target population. Although these reasons for translating tests may still apply today, cross-population comparisons are receiving the greatest amount of attention as reasons for translating tests. Certainly the recent proliferation of research on test translations is due to the interest in providing valid comparisons of traits across populations. We are increasingly viewing our world from a multicultural perspective and consequently there is a need to (1) understand the similarities and differences that exist between populations and (2) provide unbiased testing opportunities across different segments of a single population. Testing across populations provides a means for accomplishing these goals, and test translations are necessary to validly carry out this testing.

Translating test items from one language to another while attempting to maintain the original meaning of the items can be an extremely difficult task. Several potential problems associated with translating tests were noted including 1) identifying and minimizing

cultural differences, 2) identifying the appropriate language for testing target populations, 3) finding equivalent words or phrases and, 4) finding competent translators. Each of these problems alone can seriously undermine the validity of a test translation. Taken together, it easy to understand why the task of validly translating a test is a difficult one. With further research in the areas of linguistics and cross-cultural psychology, it is possible that the severity of some of these problems will generally or in certain cases be reduced. However, given the complexity of language and other cultural differences across many populations, it can be assumed that validly translating tests will continue to be a complex task.

Since there are a number of potential problems associated with translating tests, it is essential that steps be taken to insure the equivalence of a source and translated test. It was pointed out that a number of different methods for establishing the translation equivalence of tests have been used. All together, seven methods (both judgmental and statistical) of establishing translation equivalence were discussed. Six of these seven methods were identified through a review of the test translation literature. Of these seven methods, three are particularly popular (1.B.1 - source language monolinguals check for errors, 2.A.1 - bilinguals take source and target versions and 2.A.2 - source language monolinguals take the source version and target language monolinguals take the target version) while the remaining four methods have received little attention in test translations studies. The three more popular methods are used more often because they are less likely than the remaining four methods to introduce error into a test translation study.

Of the three more popular methods, method 2.A.2 is the preferred method for establishing the equivalence of translated test items. This is because method 2.A.2 does not make use of back translations (as does method 1.B.1) or bilingual examinees (as does method 2.A.1) and instead makes use of examinee samples that are similar to those who will be taking the final source and target versions of the test. This endorsement of method 2.A.2 does not mean that this method of establishing translation equivalence should be used exclusively. To the contrary, it is highly recommended that multiple methods be used if the resources for implementing them are available. Each of the seven methods of establishing translation equivalence that were discussed have unique advantages and disadvantages. By using multiple methods, the advantages of each method will accumulate, resulting in a potentially more valid test translation study.

In order to effectively use method 2.A.2, it is necessary to use a statistical technique to condition on examinee ability when comparing the scores obtained by source and target examinee samples. Of the conditional statistical techniques that are available, item response models have received the most attention in the test translation literature. The main reason for this considerable degree of interest is that within the framework of item response theory, it is possible to obtain item parameters that are independent of the specific sample of examinees used to calibrate the items. Invariant item parameters are particularly desirable in a test translation equivalence study because they provide a strong basis for taking into account differences in examinees abilities when comparing item parameters across populations. Other conditional statistical techniques that do not make use of

invariant item parameters can also be used to condition on examinee ability, but a number of problems exist with these methods making the use of item response models particularly attractive in test translation studies.

It has also been noted in the test translation literature that there is an interest in using item response models to obtain ability estimates that are not dependent on the particular items used. These invariant ability scores are useful for designing and using translated tests because they allow for placing items that will not or did not translate well on the same ability (or difficulty) scale as those that did translate well. This means that it is not necessary that all of the items in the source and target versions of a test be equivalent. As long as test items meet the assumptions of the item response model being used and measure the same trait as the items that were successfully translated, the items can be used in the population for which they were originally intended and still be used to compare examinees on the trait of interest across populations. The potential advantages of this are first, that it is possible to use tests that are more culturally relevant to each of the populations being compared and secondly, it is not necessary to attempt translating items that would be difficult to translate. Even though these benefits are not directly related to establishing translation equivalence, this is an extremely intriguing aspect of using item response theory in test translation work and will likely be a driving force in the future use of item response theory for test translation work in general.

The second and main focus of this thesis was to investigate anchor test designs when using item response theory in a translation

equivalence study. When using item response theory to establish translation equivalence, it is necessary to place the item parameters obtained in each population onto a common ability (or difficulty) scale. Corresponding test or item characteristic curves obtained from the source and target versions of a test cannot be meaningfully compared until a common metric has been established.

A confusing point concerning the scaling of these item parameters with an anchor test design is that it is not clear how many anchor items are required to provide adequate scaling. Results from the length of anchor test studies reviewed in this thesis varied, with results indicating that as few as 2 or as many as 20 anchor items are required to provide adequate scaling results. Even taking into account factors that were different across many of the studies such as scaling method, sample size, and differences in the difficulties of the tests being scaled, it is difficult to determine even an approximate appropriate anchor test length. In many testing situations this dilemma is not important since it may be relatively easy to use 20 (or more) item anchor tests. In the case of a test translation study, the number of items determined to be equivalent in the source and target populations and therefore usable as anchor items may be quite low. In these cases, determining the minimal number of anchor items that can be used and still obtain adequate scaling results becomes a more critical question. The second focus of this thesis was to answer this question under conditions similar to those that may be found in a translation equivalence study. These conditions include (1) relatively small sample sizes and (2) examinee ability overlaps that are more representative of vertical rather than horizontal scaling situations. The effect of these

two variables on the number of anchor items required to provide adequate scaling results was also investigated.

The following discussion addresses the four main questions raised in this study. The results highlighted in the discussion that follows are based on Types I and II scaling errors, but were supported by the other two methods of evaluation used in the study.

1. How do differences in calibration sample size affect the anchor test length required to provide reasonably accurate IRT scaling results? In general, larger sample sizes provide more accurate scaling results. This is because larger sample sizes result in more accurate estimation of item and ability parameters and therefore less error is introduced into the scaling process. More specifically, for the 5 and 10 item anchor tests, there was an approximately 50% reduction in the total MAD scaling error for an increase in calibration sample size from 300 to 600. For the 15 and 20 item anchor tests, an approximately 30% reduction in the MAD scaling error occurred with the same increase in sample size. Given that the amount of scaling error obtained with the N=300 samples was substantial across the four anchor test lengths, these reductions in scaling error with a doubling of the calibration sample size were certainly significant and would be helpful when conducting an IRT test translation study.

   It was also noted that comparable scaling results were obtained for the 10, 15 and 20 item anchor tests with the N=600 samples. This result suggests that for anchor tests consisting of ten or more items, anchor test length is less critical to obtaining accurate scaling results when larger sample sizes are

used. Because of this, it is recommended that larger examinee samples be used whenever possible when conducting an IRT test translation study. In general, larger examinee samples should be used to provide more accurate item and ability parameter estimates.

2. How do differences in the mean ability of examinee groups affect the anchor test length required to provide reasonably accurate IRT scaling results? In general, larger examinee ability distribution overlaps provide more accurate scaling results. This is because with larger examinee ability distribution overlaps there are more examinees located in the region of ability where the anchor items are located and therefore the parameters for these items are estimated more accurately than they would be with a smaller examinee ability overlap. More accurately estimated item parameters lead to less scaling error. More specifically, for the 5 item anchor tests, an approximately 50% reduction in the total MAD scaling error occurred with an increase in examinee ability distribution overlap from 50% to 80%. For the 10 and 15 item anchor tests, this increase in examinee ability overlap resulted in approximately a 40% reduction in the total MAD scaling error. Finally, for the 20 item anchor tests, this increase in examinee ability overlap resulted in less than a 10% reduction in the total MAD scaling error. Given that the amount of scaling error obtained with the 50% examinee ability distribution overlap was substantial across the four anchor test lengths, these reductions in scaling error with a 30% increase in ability distribution overlap were certainly significant for the 5, 10 and 15 anchor

tests and would be helpful when conducting an IRT test translation study.

It was also noted that somewhat comparable total MAD scaling error results were obtained for the 10, 15 and 20 items anchor tests with the 80% examinee ability distribution overlap samples. This result suggests that for anchor tests consisting of ten or more items, anchor test length is less critical to obtaining accurate scaling results when larger ability distribution overlap samples are used. Because of this, it is recommended that examinee samples with larger ability distribution overlaps be used whenever possible when conducting an IRT test translation study. The use of examinee samples with larger ability distribution overlaps is, of course, helpful, even when larger anchor tests are used. A priori information or possibly some type of matching variable may be useful in selecting examinee samples with larger ability distribution overlaps.

3. How does the interaction of these two factors affect the anchor test length required to provide reasonably accurate IRT scaling results? As noted previously, larger calibration sample sizes and larger examinee ability distribution overlaps both result in more accurate scaling results. Combined larger N and larger examinee ability distribution overlap samples represent ideal scaling conditions for these two variables and resulted in minimal scaling error. For example, even with an anchor test as short as 5 items, the total MAD scaling error for the N=600 and 80% ability distribution overlap sample was 0.66. This corresponds to a 1.32% error in the scaled scores for the 50 item test used in this

study. In the most extreme comparison, the total MAD scaling error for the N=300 and 50% ability distribution overlap sample was 3.89 which corresponds to a 7.78% scaling error for the 50 item test used in this study. Clearly, the combined effects of calibration sample size and examinee ability distribution overlap have a substantial effect on the accuracy of scaling results. This effect amplifies those of either the calibration sample size or examinee ability distribution overlap alone.

The previously noted results of comparable total MAD scaling error for the 10, 15 and 20 item anchor tests with the separate N=600 and the 80% examinee ability distribution overlap samples was also amplified when the results were broken down by both sample size and examinee ability distribution overlap. For the combined N=600 and 80% ability distribution overlap samples, the total MAD scaling error was extremely low for all four anchor test lengths. This result suggests that for anchor test lengths of five or more items, anchor test length is less critical for obtaining accurate scaling results when the equating design uses larger N (N=600 or higher) and larger ability distribution overlap samples (80% or higher). This result may help to explain why results concerning anchor test length in the IRT scaling literature have indicated that shorter anchor tests can provide adequate scaling results. Under ideal scaling conditions that include large sample sizes and high overlap in the distributions of examinee ability, it is not necessary to use more than a few anchor items to obtain adequate scaling results.

It was also noted that for the sample sizes and examinee ability distribution overlaps considered in this study, the reduction in total MAD scaling error with an increase in sample size for each ability distribution overlap was greater than the reduction in the total MAD scaling error with an increase in ability distribution overlap for each sample size. This appeared to be mainly due to mainly to calibration (Type 1) error alone. Based on this result, sample size appears to be more important than examinee ability distribution overlap in providing accurate item parameter estimates and therefore more accurate scaling results. This conclusion is, of course, limited in generalizability to the conditions simulated in the study.

4. What anchor test length will provide reasonably accurate IRT scaling results? Anchor test lengths of at least ten items would seem to be necessary when conducting an IRT test translation study. Longer anchor tests should be used if possible particularly with smaller sample sizes and suspected smaller examinee ability distribution overlaps, but a minimum of ten item anchor tests will provide fairly comparable scaling results to those obtained with longer anchor tests even with relatively poor scaling designs (such as the N=300 and 50% examinee ability distribution overlap sample design used in this study). It is also reasonable in a test translation study to use an anchor test consisting of ten well translated items rather than a longer anchor test that contains items of questionable translation equivalence.

It is possible to obtain relatively good scaling results with anchor test lengths as short as 5 items under ideal scaling designs that include larger sample sizes and larger examinee ability distribution overlaps (such as the N=600 and 80% examinee ability distribution overlap sample design used in this study). Even though it is possible to obtain relatively good scaling results with this length anchor test, it is not advisable to use a 5 item anchor test in a test translation study unless additional anchor items can not be found. This is because the use of even one poor anchor item can have a substantial negative effect on the scaling results when an anchor test of this length is used. In addition, it is unrealistic to expect large examinee ability distribution overlaps for the samples used in many test translation studies even if large samples can be used. Since a five item anchor tests generally resulted in substantial scaling error, it is recommended that a minimum of 10 item anchor tests consisting of well translated items be used when conducting a translation equivalence study. If additional items of established translation equivalence are available, longer anchor tests should be used.

A few comments on the generalizability of the scaling error results obtained in this study are in order. These scaling error results are likely to be lower than those that would be obtained in actual test translation studies conducted under similar conditions to those used in this study. There are two reasons for this. First, problems with model-data fit were not encountered. The simulated data used in this study were generated using a three parameter logistic model

with pseudo-chance (c) parameters set to 0.2 and, in addition, the data were unidimensional. Since a unidimensional three parameter logistic model with a pseudo-chance parameter fixed at 0.2 was used to calibrate the items for the simulated data, problems related to model-data misfit were not encountered. Under conditions of an actual test translation study, some degree of model-data misfit would be expected and poorer scaling results than those obtained in this study would likely result.

Second, even though few aberrant item difficulty estimates (greater or less than 4.0) were obtained during item calibration, data sets containing these aberrant results were not used. Consequently, extreme item difficulty estimates did not effect the results obtained in this study. In practice, aberrant item difficulty estimates may occur, and they can adversely effect (1) scaling results if they are obtained for items used in an anchor test or (2) item translation equivalence results if they are obtained for unique (non-anchor) items.

Also, it is not clear what effects fixing the pseudo-chance (c) item parameter at a specific value had on the generalizability of the scaling results obtained in this study. The pseudo-chance item parameter is often the most difficult item parameter to estimate because the number of examinees at the lower end of the examinee ability distribution is typically small. By fixing the pseudo-chance parameter at a specific value during the item calibration, problems with estimating this item parameter are eliminated. Also, problems with estimating item and ability parameters in general are reduced. This possibly allows for the use of smaller sample sizes and reduces the cost of performing IRT computer simulation studies. Since it is unrealistic in many testing situations (including test translation studies) to assume that no

guessing or minimal guessing at items occurs, fixing the pseudo-chance item parameter at a specific value allows for taking guessing at items into account while at the same time reducing parameter estimation problems. However, it is not clear how fixing the pseudo-chance parameter would effect the results of many test translation studies and to the extent that this effect is not known, the generalizability of the results obtained in this study is reduced somewhat.

In conclusion, several suggestions for additional test translation research are offered:

1. Research into methods of insuring the equivalence of anchor items. Identifying anchor items that are actually equivalent in source and target populations is critical to the validity of any IRT test translation study. The apparent circularity of attempting to place item parameters for the source and target versions of a test on the same scale by using anchor items identified through assuming a common scale is problematic. Criterion purification procedures have been developed to limit this problem, but further research into more refined procedures is warranted.
2. Research into the use of anchor items developed by using source language items that are relatively easy to translate. Given the importance and problems of identifying anchor items in a test translation study, it may be helpful to develop test items in the source language specifically so they can be translated easily and meaningfully for use as anchor items. These items could then be cycled through the normal procedures for identifying anchor items so their status as anchor items could be empirically confirmed.

Studies on this method of helping to obtain anchor tests would be an extremely useful addition to test translation research.

3. Research into the problems of effectively translating test items. Most test translation research or descriptions of test translation projects or research focus on either the methodology used to establish translation equivalence and/or the final translation equivalence results. Little is mentioned about the difficulties of translating specific items beyond perhaps providing a few narrative examples. Research that focuses on the problems of test translations at the item level might be helpful to those attempting test translations by highlighting potential pitfalls and possible explanations for the non-equivalence of translated items. A substantial amount of in depth research into this area could also provide general rules or guidelines concerning the types (i.e., content, item format, etc.) of items that are problematic in test translations. These problematic item types could be avoided or given special attention when translating tests.

4. Research into the effects of different types of model-data misfit on the results of an IRT test translation study. For example, item parameter invariance is particularly important when conducting an IRT test translation study and this expected model feature can only be obtained when there is an adequate fit of an IRT model to the data being used. Therefore, research into the effects of violations of different types of IRT model assumptions on item parameter invariance would be particularly useful. Since it is unlikely that a particular IRT model will fit the data from

source and target populations in the same way, an understanding of the robustness of IRT models to violations of their assumptions is especially pertinent.

APPENDIX A

TRANSLATED TESTS AND QUESTIONNAIRES/INVENTORIES

Translated Tests and Questionnaires/Inventories

A. Tests

Name	Source Language	Target Language
Bennet Mechanical Comprehension Test	English	Spanish
College Board Scholastic Aptitude Test	English	Spanish
Differential Aptitude Test	English	Spanish
Inter-American Series Test of General Ability	English	Spanish
Stanford-Binet Intelligence Scale	English	Spanish
Wechsler Intelligence Scale for Children	English	Spanish
Western Personnel Test	English	Spanish

B. Questionnaires/Inventories

Name	Source Language	Target Language
Association Adjustment Survey	English	Spanish
California Occupational Preference Survey	English	Spanish
Curtis Completion Form	English	Spanish
Index of Organizational Reactions	English	Spanish
Job Descriptive Index	English	Spanish-Hebrew
Strong-Campbell Interest Inventory	English	Spanish
STS Youth Inventory (Form G)	English	Spanish
Vocational Preference Inventory	English	Spanish

APPENDIX B

COMPUTER PROGRAMS FOR ESTIMATING ITEM RESPONSE MODEL ITEM AND ABILITY PARAMETERS[1]

[1]Adapted from Hambleton (1979)

Name	Model Applications	Estimation Procedure	Computing Environment
BICAL	one-parameter logistic	joint maximum likelihood	Mainframe
LOGIST	one-, two-, or three-parameter logistic model	joint maximum likelihood	Mainframe
BILOG	one-, two-, or three-parameter logistic model	marginal maximum likelihood with optimal Bayesian priors	Mainframe, IBM PC, XT, or AT
MicroCAT	one-, two-, or three-parameter logistic model	combined maximum likelihood and Bayesian	IBM PC, XT, or AT
ANCILLES: OGIVA	three-parameter logistic model	Urry estimation procedure	Mainframe

APPENDIX C

PROGRAM 1

```
      PROGRAM SCALE(TAPE5,TAPE7,TAPE8,TAPE17,TAPE18,TAPE21)
C
C     ****************************************
C     THIS PROGRAM SCALES THE TRUE SCORES FOR
C     TWO GROUPS USING THE CHARACTERISTIC
C     CURVE SCALING METHOD
C     ****************************************
C
      REAL ABLT(2,1000),A(2,100),B(2,100),C(2,100)
      REAL SUMAB(2),SIGMAB(2),AVGAB(2),SDAB(2)
      REAL SUMB(2),SUMA(2),SUMC(2),MAXB(2),MAXA(2),MAXC(2)
      REAL AVGB(2),MINB(2),AVGA(2),MINA(2),AVGC(2),MINC(2)
      REAL ASUMB(2),ASUMA(2),ASUMC(2),AMAXB(2),AMAXA(2),AMAXC(2)
      REAL AAVGB(2),AMINB(2),AAVGA(2),AMINA(2),AAVGC(2),AMINC(2)
      REAL UTRSC(2,1000),ATRSC(2,1000),TATRSC(1000),PTAL,PTBE
      REAL USUMT(2),UAVGT(2),USIGT(2),USDT(2),INC
      REAL PEE,DB(50),DA(50),POW,PR,SE
      REAL ASUMT(2),AAVGT(2),ASIGT(2),ASDT(2)
      REAL ALPHAT,BETAT
      REAL PPB,PT,PPA,SUMFAL,SUMFBE,X(2),PFAL(1000),PFBE(1000)
      REAL SUMDFB,SUMDFA,SUMDFB2(2),SUMDFA2(2),SUM(1000),F(1000)
      REAL SDB(2),SDA(2),POW1,POW2,PROB1,PROB2
      REAL PFAL2(1000),PFBE2(1000),SUM2(500),X1(100),X2(100)
      DIMENSION TTRSCD(-99:99),TTRSCT(-99:99),TTRSCF(-99:99)
      INTEGER NSUBJ,NITEMS,NAITEMS,REPL,AO,G,HNSUBJ,TRIAL
      INTEGER NUITEMS,SAITEMS,CUITEMS,AUITEMS,CAITEMS,AAITEMS
      CHARACTER*1 TITLE(80)
      CHARACTER*10 NAMEA,NAMEB
C
C     TAPE5=INPUT PARAMETERS:
C
C       LINE1 TITLE (80A1)
C
C       LINE2 GROUP NAMES, RUN NUMBER (DESIGNATES REPLICATION),
C             ABILITY OVERLAP (2A10,2I5)
C
C       LINE3 NUMBER OF EXAMINEES, TOTAL ITEMS, ANCHOR ITEMS (3I5)
C
C     TAPE7=ABILITY AND ITEM PARAMETERS FOR GROUP 1 (LOW ABILITY)
C     TAPE8=ABILITY AND ITEM PARAMETERS FOR GROUP 2 (HIGH ABILITY)
C
      READ (5,5) TITLE
    5 FORMAT (80A1)
C
      READ (5,10) NAMEA,NAMEB,REPL,AO
   10 FORMAT (2A10,2I5)
C
      READ (5,15) NSUBJ,NITEMS,NAITEMS
   15 FORMAT(3I5)
      HNSUBJ=NSUBJ/2
C
      WRITE(REPL,20) TITLE,NAMEA,NAMEB,HNSUBJ,NITEMS,NAITEMS,AO
   20 FORMAT(//15X,80A1/15X,A10,3X,A10/15X,'SAMPLE SIZE/GROUP =',I5/
```

```
     115X,'TOTAL NUMBER OF ITEMS -',I5/15X,'NUMBER OF ANCHOR ITEMS -'
     2,I5/15X,'ABILITY OVERLAP -',I5,'%')
C
      READ(7,25)
   25 FORMAT(/////)
      READ(7,30) (ABLT(1,J),J-1,HNSUBJ)
   30 FORMAT(4X,9F10.3/10F10.3)
C
      READ(8,25)
      READ(8,30) (ABLT(2,J),J-1,HNSUBJ)
C
      IF(AO.EQ.50) THEN
      ALPHAT-1.0
      BETAT-1.35
      ELSE IF (AO.EQ.80) THEN
      ALPHAT-1.0
      BETAT-0.51
      ELSE
      ALPHAT-1.0
      BETAT-0.0
      ENDIF
C
C
      NN-NITEMS/6
      NNN-NN*6
      IF (NNN.NE.NITEMS) NN-NN+1
      DO 1010 J-0,NN
      M1-1+6*J
      M2-6*(J+1)
      IF (M2.GT.NITEMS) M2-NITEMS
      READ(7,40) (A(1,I),B(1,I),C(1,I),I-M1,M2)
   40 FORMAT (4X,9F10.3/9F10.3)
 1010 CONTINUE
      DO 2010 J-0,NN
      M1-1+6*J
      M2-6*(J+1)
      IF (M2.GT.NITEMS) M2-NITEMS
      READ(8,40) (A(2,I),B(2,I),C(2,I),I-M1,M2)
 2010 CONTINUE
C
C     ...FLAG ITEMS WITH POORLY ESTIMATED B'S...
C
      NUITEMS-NITEMS-NAITEMS
      SAITEMS-NUITEMS+1
      CUITEMS-0
      CAITEMS-0
      DO 1012 G-1,2
      DO 1014 I-1,NUITEMS
      IF(B(2,I).EQ.99.0) GOTO 52
      IF(ABS(B(G,I)).GT.4.0) THEN
      B(1,I)-99.0
      B(2,I)-99.0
      CUITEMS-CUITEMS+1
   52 ENDIF
```

```
 1014 CONTINUE
 1012 CONTINUE
      DO 1016 G=1,2
      DO 1018 I=SAITEMS,NITEMS
      IF(B(2,I).EQ.99.0) GOTO 54
      IF(ABS(B(G,I)).GT.4.0) THEN
      B(1,I)=99.0
      B(2,I)=99.0
      CAITEMS=CAITEMS+1
   54 ENDIF
 1018 CONTINUE
 1016 CONTINUE
      AUITEMS=NUITEMS-CUITEMS
      AAITEMS=NAITEMS-CAITEMS
      PRINT*,AUITEMS,'      ',AAITEMS
C
C
C     ...CALCULATE MEANS AND STANDARD DEVIATIONS
C        OF EXAMINEE ABILITY FOR EACH GROUP...
C
      DO 1025 G=1,2
      SUMAB(G)=0.0
      DO 1030 J=1,HNSUBJ
      IF(ABLT(G,J).GT.3.00) ABLT(G,J)=3.0
      IF(ABLT(G,J).LT.-3.00) ABLT(G,J)=-3.0
      SUMAB(G)=SUMAB(G)+ABLT(G,J)
 1030 CONTINUE
 1025 CONTINUE
      DO 1035 G=1,2
      SIGMAB(G)=0.0
      AVGAB(G)=SUMAB(G)/FLOAT(HNSUBJ)
      DO 1040 J=1,HNSUBJ
      SIGMAB(G)=SIGMAB(G)+(ABLT(G,J)-AVGAB(G))**2
 1040 CONTINUE
      SDAB(G)=SQRT(SIGMAB(G)/(FLOAT(HNSUBJ)-1.0))
 1035 CONTINUE
C
C     ...CALCULATE MEANS AND MIN/MAX OF UNIQUE
C        ITEM PARAMETERS FOR EACH TEST...
C
      DO 1045 G=1,2
      SUMB(G)=0.0
      SUMA(G)=0.0
      SUMC(G)=0.0
      DO 1050 I=1,NUITEMS
      IF(B(G,I).EQ.99.0) GOTO 1050
      SUMB(G)=SUMB(G)+B(G,I)
      SUMA(G)=SUMA(G)+A(G,I)
      SUMC(G)=SUMC(G)+C(G,I)
 1050 CONTINUE
 1045 CONTINUE
      DO 1055 G=1,2
      MAXB(G)=0.0
      MAXA(G)=0.0
```

```
      MAXC(G)=0.2
      MINB(G)=0.0
      MINA(G)=0.0
      MINC(G)=0.2
      AVGB(G)=SUMB(G)/FLOAT(AUITEMS)
      AVGA(G)=SUMA(G)/FLOAT(AUITEMS)
      AVGC(G)=SUMC(G)/FLOAT(AUITEMS)
      DO 1060 I=1,NUITEMS
      IF(B(G,I).EQ.99.0) GOTO 1060
      IF(B(G,I).GT.MAXB(G)) MAXB(G)=B(G,I)
      IF(B(G,I).LT.MINB(G)) MINB(G)=B(G,I)
      IF(A(G,I).GT.MAXA(G)) MAXA(G)=A(G,I)
      IF(A(G,I).LT.MINA(G)) MINA(G)=A(G,I)
 1060 CONTINUE
 1055 CONTINUE
C
C     ...CALCULATE MEANS AND MIN/MAX OF ANCHOR
C        ITEM PARAMETERS FOR EACH TEST...
C
      DO 1065 G=1,2
      ASUMB(G)=0.0
      ASUMA(G)=0.0
      ASUMC(G)=0.0
      DO 1070 I=SAITEMS,NITEMS
      IF(B(G,I).EQ.99.0) GOTO 1070
      ASUMB(G)=ASUMB(G)+B(G,I)
      ASUMA(G)=ASUMA(G)+A(G,I)
      ASUMC(G)=ASUMC(G)+C(G,I)
 1070 CONTINUE
 1065 CONTINUE
      DO 1075 G=1,2
      SUMDFB2(G)=0.0
      SUMDFA2(G)=0.0
      SDB(G)=0.0
      SDA(G)=0.0
      AMAXB(G)=0.0
      AMAXA(G)=0.0
      AMAXC(G)=0.2
      AMINB(G)=0.0
      AMINA(G)=0.0
      AMINC(G)=0.2
      AAVGB(G)=ASUMB(G)/FLOAT(AAITEMS)
      AAVGA(G)=ASUMA(G)/FLOAT(AAITEMS)
      AAVGC(G)=ASUMC(G)/FLOAT(AAITEMS)
      DO 1072 I=SAITEMS,NITEMS
      IF(B(G,I).EQ.99.0) GOTO 1072
      SUMDFB2(G)=SUMDFB2(G)+(B(G,I)-AAVGB(G))**2
      SUMDFA2(G)=SUMDFA2(G)+(A(G,I)-AAVGA(G))**2
      IF(B(G,I).GT.AMAXB(G)) AMAXB(G)=B(G,I)
      IF(B(G,I).LT.AMINB(G)) AMINB(G)=B(G,I)
      IF(A(G,I).GT.AMAXA(G)) AMAXA(G)=A(G,I)
      IF(A(G,I).LT.AMINA(G)) AMINA(G)=A(G,I)
 1072 CONTINUE
      SDB(G)=SQRT(SUMDFB2(G)/(FLOAT(AAITEMS)-1.0))
```

```
      SDA(G)=SQRT(SUMDFA2(G)/(FLOAT(AAITEMS)-1.0))
 1075 CONTINUE
C
C     ...CORRELATIONS OF ANCHOR ITEM PARAMETERS
C        FOR TESTS 1 AND 2...
C
      DO 1078 I=SAITEMS,NITEMS
      DIFF1=0.0
      DIFF2=0.0
      DIFF12=0.0
      IF((B(1,I).EQ.99.0) .OR. (B(2,I).EQ.99.0)) GOTO 1078
      DIFF1=B(1,I)-AAVGB(1)
      DIFF2=B(2,I)-AAVGB(2)
      DIFF12=DIFF1*DIFF2
      SUMDFB=SUMDFB+DIFF12
 1078 CONTINUE
      COVB=SUMDFB/FLOAT(AAITEMS-1)
      DO 1080 I=SAITEMS,NITEMS
      DIFFA1=0.0
      DIFFA2=0.0
      DIFFA12=0.0
      IF((B(1,I).EQ.99.0) .OR. (B(2,I).EQ.99.0)) GOTO 1080
      DIFFA1=A(1,I)-AAVGA(1)
      DIFFA2=A(2,I)-AAVGA(2)
      DIFFA12=DIFFA1*DIFFA2
      SUMDFA=SUMDFA+DIFFA12
 1080 CONTINUE
      COVA=SUMDFA/FLOAT(AAITEMS-1)
      CORRB=COVB/(SDB(1)*SDB(2))
      CORRA=COVA/(SDA(1)*SDA(2))
C
      WRITE(REPL,75) HNSUBJ,AVGAB(1),SDAB(1),AVGAB(2),SDAB(2)
   75 FORMAT(////14X,'SUMMARY STATISTICS OF EXAMINEE ABILITY
     1(N=',I5,')'//13X,50('-')/16X,'GROUP',14X,'MEAN',14X,
     2'SD'/13X,50('-')//18X,'1',15X,F4.2,14X,F4.2/18X,'2',15X,
     3F4.2,14X,F4.2//13X,50('-'))
      WRITE(REPL,85) NUITEMS,AVGB(1),MAXB(1),AVGA(1),MAXA(1),
     1AVGC(1),MAXC(1),MINB(1),MINA(1),MINC(1),AVGB(2),MAXB(2),
     2AVGA(2),MAXA(2),AVGC(2),MAXC(2),MINB(2),MINA(2),
     3MINC(2),AUITEMS
   85 FORMAT(//9X,'SUMMARY STATISTICS OF UNIQUE ITEM
     1PARAMETERS (N=',I5,')'/5X,68('-')/27X,'B',17X,'A',17X,
     2'C'/5X,68('-')/8X,'GROUP',8X,'MEAN',5X,'MAX/',5X,'MEAN',
     35X,'MAX/',5X,'MEAN',5X,'MAX/'/30X,'MIN',15X,'MIN',
     415X,'MIN'/5X,68('-')//
     510X,'1',9X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2/
     628X,F5.2,13X,F5.2,13X,F5.2//
     710X,'2',9X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2/
     828X,F5.2,13X,F5.2,13X,F5.2//
     95X,68('-')/14X,'ACTUAL NUMBER OF UNIQUE ITEMS= ',I3)
      WRITE(REPL,90) NAITEMS,AAVGB(1),AMAXB(1),AAVGA(1),
     1AMAXA(1),AAVGC(1),AMAXC(1),AMINB(1),AMINA(1),AMINC(1),
     2AAVGB(2),AMAXB(2),AAVGA(2),AMAXA(2),AVGC(2),AMAXC(2),
     3AMINB(2),AMINA(2),AMINC(2),AAITEMS
```

```
   90 FORMAT(//9X,'SUMMARY STATISTICS OF ANCHOR ITEM
      PARAMETERS (N=',I5,')'/5X,68('-')/27X,'B',17X,'A',17X,
     1'C'/5X,68('-')/8X,'GROUP',8X,'MEAN',5X,'MAX/',5X,
     2'MEAN',5X,'MAX/',5X,'MEAN',5X,'MAX/'/30X,'MIN',15X,
     3'MIN',15X,'MIN'/5X,68('-')//10X,'1',9X,F5.2,3X,F5.2,
     45X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2/28X,F5.2,13X,F5.2,13X,
     5F5.2//10X,'2',9X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2,5X,
     6F5.2,3X,F5.2/28X,F5.2,13X,F5.2,13X,F5.2//
     75X,68('-')/14X,'ACTUAL NUMBER OF ANCHOR ITEMS= ',I3)
C
C    ...CALCULATE ITEM CHARACTERISTIC CURVES/TRUE SCORES
C       FOR UNIQUE ITEMS...
C
      DO 1085 G=1,2
      DO 1090 J=1,HNSUBJ
      UTRSC(G,J)=0.0
      DO 1095 I=1,NUITEMS
      IF (B(G,I).EQ.99.0) GOTO 1095
      POW=(1.7*(ABLT(G,J)-B(G,I))*A(G,I))
      PR=C(G,I)+(1.0-C(G,I))*EXP(POW)/(1.+EXP(POW))
      UTRSC(G,J)=UTRSC(G,J)+PR
 1095 CONTINUE
 1090 CONTINUE
 1085 CONTINUE
C
C    ...CALCULATE ITEM CHARACTERISTIC CURVES/TRUE SCORES
C       FOR ANCHOR ITEMS...
C
      DO 1100 G=1,2
      DO 1105 J=1,HNSUBJ
      ATRSC(G,J)=0.0
      DO 1110 I=SAITEMS,NITEMS
      IF (B(G,I).EQ.99.0) GOTO 1110
      POW=(1.7*(ABLT(G,J)-B(G,I))*A(G,I))
      PR=C(G,I)+(1.0-C(G,I))*EXP(POW)/(1.+EXP(POW))
      ATRSC(G,J)=ATRSC(G,J)+PR
 1110 CONTINUE
 1105 CONTINUE
 1100 CONTINUE
      DO 1230 G=1,2
      DO 1150 J=1,HNSUBJ
 1150 CONTINUE
 1230 CONTINUE
C
C     ...CALCULATE MEANS AND STANDARD DEVIATIONS OF
C        TRUE SCORES ON UNIQUE ITEMS...
C
      DO 1160 G=1,2
      USUMT(G)=0.0
      DO 1165 J=1,HNSUBJ
      USUMT(G)=USUMT(G)+UTRSC(G,J)
 1165 CONTINUE
 1160 CONTINUE
      DO 1170 G=1,2
```

```
      USIGT(G)=0.0
      USDT(G)=0.0
      UAVGT(G)=USUMT(G)/FLOAT(HNSUBJ)
      DO 1175 J=1,HNSUBJ
      USIGT(G)=USIGT(G)+(UTRSC(G,J)-UAVGT(G))**2
 1175 CONTINUE
      USDT(G)=SQRT(USIGT(G)/(FLOAT(HNSUBJ)-1.0))
 1170 CONTINUE
C
C     ...CALCULATE MEANS AND STANDARD DEVIATIONS OF
C        TRUE SCORES ON ANCHOR ITEMS...
C
      DO 1180 G=1,2
      ASUMT(G)=0.0
      DO 1185 J=1,HNSUBJ
      ASUMT(G)=ASUMT(G)+ATRSC(G,J)
 1185 CONTINUE
 1180 CONTINUE
      DO 1190 G=1,2
      ASIGT(G)=0.0
      AAVGT(G)=ASUMT(G)/FLOAT(HNSUBJ)
      DO 1195 J=1,HNSUBJ
      ASIGT(G)=ASIGT(G)+(ATRSC(G,J)-AAVGT(G))**2
 1195 CONTINUE
      ASDT(G)=SQRT(ASIGT(G)/(FLOAT(HNSUBJ)-1.0))
 1190 CONTINUE
C
      WRITE(REPL,95) UAVGT(1),USDT(1),AAVGT(1),ASDT(1),
     1UAVGT(2),USDT(2),AAVGT(2),ASDT(2)
   95 FORMAT(//9X,'SUMMARY STATISTICS OF TRUE SCORES'/
     15X,68('-')/25X,'UNIQUE ITEMS',12X,'ANCHOR ITEMS'/
     25X,68('-')/10X,'GROUP',11X,'MEAN',3X,'SD',15X,
     3'MEAN',3X,'SD'/5X,68('-')//12X,'1',11X,2F6.2,12X,
     32F6.2//12X,'2',11X,2F6.2,12X,2F6.2//
     45X,68('-'))
C
      PRINT *,'ENTER INITIAL VALUE FOR ALPHA:'
      READ*,X(1)
      PRINT *,'ENTER INITIAL VALUE FOR BETA:'
      READ *,X(2)
C
      WRITE(REPL,100) NSUBJ,NITEMS,NAITEMS,REPL,AO,X(1),
     1X(2),CORRB,CORRA
  100 FORMAT(//5I5/4F10.3)
      DO 1225 G=1,2
      DO 1228 J=1,HNSUBJ
      WRITE(REPL,105) ABLT(G,J)
  105 FORMAT(F12.3)
 1228 CONTINUE
 1225 CONTINUE
      DO 1235 G=1,2
      WRITE(REPL,110) AVGB(G),AVGA(G),AVGC(G),AAVGB(G),
     1AAVGA(G),AAVGC(G)
  110 FORMAT(6F6.3)
```

```
 1235 CONTINUE
      DO 1240 G=1,2
      WRITE(REPL,115) MAXB(G),MAXA(G),AMAXB(G),AMAXA(G),
     1MINB(G),MINA(G),AMINB(G),AMINA(G)
  115 FORMAT(8F10.3)
 1240 CONTINUE
C
C     ...CALCULATE SCALING COEFFICIENTS...
C
      TRIAL=1
      L=1
      K=0
 1270 SUMFAL=0.0
      SUMFBE=0.0
      INC=.01
      DO 1272 J=1,HNSUBJ
C
C      ...GET TRUE SCORES FOR EXAMINEES IN GROUP 1
C         USING EACH SET OF ITEM PARAMETERS...
C
      TRSCA=0.0
      TRSCB=0.0
      DO 1274 I=SAITEMS,NITEMS
      IF((B(1,I).EQ.99.0) .OR. (B(2,I).EQ.99.0)) GOTO 1274
      ATRAN=A(2,I)/X(1)
      BTRAN=B(2,I)*X(1)+X(2)
      CTRAN=C(2,I)
      POW1=1.7*(ABLT(1,J)-B(1,I))*A(1,I)
      EX1=EXP(POW1)
      PROB1=C(1,I)+(1.0-C(1,I))*EXP(POW1)/(1.+EXP(POW1))
      TRSCA=TRSCA+PROB1
      POW2=1.7*(ABLT(1,J)-BTRAN)*ATRAN
      EX2=EXP(POW2)
      PROB2=CTRAN+(1.0-CTRAN)*EXP(POW2)/(1.0+EXP(POW2))
      TRSCB=TRSCB+PROB2
 1274 CONTINUE
      PTAL=0.0
      PTBE=0.0
      TDIFF=TRSCA-TRSCB
      TDIFF2=(TRSCA-TRSCB)**2
C
C     ...GET FIRST DERIVATIVES OF F WRT ALPHA AND BETA...
C
      DO 1276 I=SAITEMS,NITEMS
      IF((B(1,I).EQ.99.0) .OR. (B(2,I).EQ.99.0)) GOTO 1276
      ATRAN=A(2,I)/X(1)
      BTRAN=B(2,I)*X(1)+X(2)
      CTRAN=C(2,I)
      PT=1.7*(ABLT(1,J)-BTRAN)*(1.0-PROB2)*(PROB2-CTRAN)/
     1(1.0-CTRAN)
      PPB=-1.7*ATRAN*(1.0-PROB2)*(PROB2-CTRAN)/(1.0-CTRAN)
      PPA=B(2,I)*PPB-A(2,I)*PT/SQRT(X(1))
      PTAL=PTAL+PPA
      PTBE=PTBE+PPB
```

```
 1276 CONTINUE
      SUMF=SUMF+TDIFF2
      SUMFAL=SUMFAL+TDIFF*PTAL
      SUMFBE=SUMFBE+TDIFF*PTBE
 1272 CONTINUE
C
C      ...PFAL(L) IS THE PARTIAL OF F WRT ALPHA...
C      ...PFBE(L) IS THE PARTIAL OF F WRT BETA...
C
      F(L)=SUMF/FLOAT(HNSUBJ)
      PFAL(L)=(-2.0/FLOAT(HNSUBJ))*SUMFAL
      PFBE(L)=(-2.0/FLOAT(HNSUBJ))*SUMFBE
C
 1275 IF(L.EQ.1) GOTO 1280
      SUM(L)=ABS(PFAL(L))+ABS(PFBE(L))
      IF(TRIAL.EQ.2) GOTO 1285
      IF(TRIAL.EQ.3) GOTO 1295
      IF(TRIAL.EQ.4) GOTO 1305
      IF(TRIAL.EQ.5) GOTO 1315
      IF(TRIAL.EQ.6) GOTO 1325
      IF(TRIAL.EQ.7) GOTO 1335
      IF(TRIAL.EQ.8) GOTO 1345
      IF(SUM(L-1).LE.SUM(L)) THEN
      X(1)=X(1)-INC
      X(2)=X(2)-INC
      GOTO 1290
      ELSE
      GOTO 1280
      ENDIF
 1280 X(1)=X(1)+INC
      X(2)=X(2)+INC
      L=L+1
      GOTO 1270
 1285 IF(SUM(L-1).LE.SUM(L)) THEN
      X(1)=X(1)+INC
      X(2)=X(2)+INC
      GOTO 1300
      ELSE
      GOTO 1290
      ENDIF
 1290 TRIAL=2
      X(1)=X(1)-INC
      X(2)=X(2)-INC
      L=L+1
      GOTO 1270
 1295 IF(SUM(L-1).LE.SUM(L)) THEN
      X(1)=X(1)-INC
      X(2)=X(2)+INC
      GOTO 1310
      ELSE
      GOTO 1300
      ENDIF
 1300 TRIAL=3
      X(1)=X(1)+INC
```

```
      X(2)=X(2)-INC
      L=L+1
      GOTO 1270
 1305 IF(SUM(L-1).LE.SUM(L)) THEN
      X(1)=X(1)+INC
      X(2)=X(2)-INC
      GOTO 1320
      ELSE
      GOTO 1310
      ENDIF
 1310 TRIAL=4
      X(1)=X(1)-INC
      X(2)=X(2)+INC
      L=L+1
      GOTO 1270
 1315 IF(ABS(PFAL(L-1)).LT.ABS(PFAL(L)) .AND.
     1SUM(L-1).LT.SUM(L) .OR.
     1(SUM(L).LT.0.01)) THEN
      X(1)=X(1)-INC
      K=K+1
      PRINT*,'K= ',K
      GOTO 1330
      ELSE
      GOTO 1320
      ENDIF
 1320 IF(SUM(L).LT.0.01) GOTO 1420
      TRIAL=5
      X(1)=X(1)+INC
      L=L+1
      GOTO 1270
 1325 IF(ABS(PFAL(L-1)).LT.ABS(PFAL(L)) .AND.
     1SUM(L-1).LT.SUM(L) .OR.
     1(SUM(L).LT.0.01)) THEN
      X(1)=X(1)+INC
      GOTO 1340
      ELSE
      GOTO 1330
      ENDIF
 1330 IF(SUM(L).LT.0.01) GOTO 1420
      TRIAL=6
      X(1)=X(1)-INC
      L=L+1
      GOTO 1270
 1335 IF(ABS(PFBE(L-1)).LT.ABS(PFBE(L)) .AND.
     1SUM(L-1).LT.SUM(L) .OR.
     1(SUM(L).LT.0.01)) THEN
      X(2)=X(2)-INC
      GOTO 1350
      ELSE
      GOTO 1340
      ENDIF
 1340 IF(SUM(L).LT.0.01) GOTO 1420
      TRIAL=7
      X(2)=X(2)+INC
```

```
      L=L+1
      GOTO 1270
 1345 IF(ABS(PFBE(L-1)).LT.ABS(PFBE(L)) .AND.
     1SUM(L-1).LT.SUM(L) .OR.
     1(SUM(L).LT.0.01)) THEN
      X(2)=X(2)+INC
      GOTO 1360
      ELSE
      GOTO 1350
      ENDIF
 1350 IF(SUM(L).LT.0.01) GOTO 1420
      TRIAL=8
      X(2)=X(2)-INC
      L=L+1
      GOTO 1270
C
 1360 IF(K.EQ.7) THEN
      GOTO 1420
      ELSE
      GOTO 1320
      ENDIF
C
 1420 WRITE(REPL,120) X(1),X(2),ALPHAT,BETAT,F(L-1)
  120 FORMAT(2F10.3/3F10.3)
      DO 1425 G=1,2
      DO 1430 J=1,HNSUBJ
      WRITE(REPL,122) UTRSC(G,J)
  122 FORMAT(F6.3)
 1430 CONTINUE
 1425 CONTINUE
      DO 1435 G=1,2
      DO 1440 J=1,HNSUBJ
      WRITE(REPL,124) ATRSC(G,J)
  124 FORMAT(F6.3)
 1440 CONTINUE
 1435 CONTINUE
C
      WRITE(17,133)
  133 FORMAT(3X,'ABL',4X,'TABT',3X,'DTRSC2',2X,'ETRSC2',3X,
     1'TTRSCD',2X,'TTRSCT',2X,'TTRSCF',6X,'PEE',3X,
     2'SE'/1X,73('-'))
C
      BIT=0.1
      DO 1443 I=1,NUITEMS
      READ(13,126) DA(I),DB(I)
  126 FORMAT(2F10.3)
      DA(I)=DA(I)*ALPHAT
      DB(I)=DB(I)/ALPHAT-BETAT
 1443 CONTINUE
      DO 1445 ABL=-3.0,3.0, BIT
      TTRSCD(ABL)=0.0
      TTRSCT(ABL)=0.0
      TTRSCF(ABL)=0.0
      DTRSC2=0.0
```

```
      ETRSC2=0.0
      PEE=0.0
      SE=0.0
      TABT=ABL*ALPHAT+BETAT
      DO 1450 I=1,NUITEMS
      IF((B(1,I).EQ.99.0) .OR. (B(2,I).EQ.99.0)) GOTO 1450
C
      ATRANF=A(2,I)/X(1)
      BTRANF=B(2,I)*X(1)+X(2)
      POW=1.7*(TABT-BTRANF)*ATRANF
      PR=0.2+0.8*EXP(POW)/(1.0+EXP(POW))
      TTRSCF(ABL)=TTRSCF(ABL)+PR
C
      ATRANT=A(2,I)/ALPHAT
      BTRANT=B(2,I)*ALPHAT+BETAT
      POW=1.7*(TABT-BTRANT)*ATRANT
      PR=0.2+0.8*EXP(POW)/(1.0+EXP(POW))
      TTRSCT(ABL)=TTRSCT(ABL)+PR
C
      POW=1.7*(ABL-DB(I))*DA(I)
      PR=0.2+0.8*EXP(POW)/(1.0+EXP(POW))
      DTRSC2=DTRSC2+PR
C
      POW=1.7*(ABL-B(2,I))*A(2,I)
      PR=0.2+0.8*EXP(POW)/(1.0+EXP(POW))
      ETRSC2=ETRSC2+PR
C
      ATRAND=DA(I)/ALPHAT
      BTRAND=DB(I)*ALPHAT+BETAT
      POW=1.7*(TABT-BTRAND)*ATRAND
      PR=0.2+0.8*EXP(POW)/(1.0+EXP(POW))
      TTRSCD(ABL)=TTRSCD(ABL)+PR
 1450 CONTINUE
      PEE=TTRSCD(ABL)-TTRSCT(ABL)
      SE=TTRSCT(ABL)-TTRSCF(ABL)
      WRITE(17,135) ABL,TABT,DTRSC2,ETRSC2,TTRSCD(ABL),TTRSCT(ABL),
     1TTRSCF(ABL),PEE,SE
  135 FORMAT(2X,F5.2,2X,F5.2,3X,F5.2,3X,F5.2,4X,F5.2,3X,F5.2,
     13X,F5.2,5X,F5.2,2X,F5.2)
      WRITE(REPL,140)TABT,TTRSCD(ABL),TTRSCT(ABL),TTRSCF(ABL)
  140 FORMAT(4F10.3)
 1445 CONTINUE
C
      DO 1455 I=1,NUITEMS
      WRITE(REPL,145)B(2,I),A(2,I)
  145 FORMAT(2F10.2)
      WRITE(21,190) B(2,I),DB(I),A(2,I),DA(I)
  190 FORMAT(//'B(2,I)= ',F7.2,'  DB(I)= ',F7.2,
     1'      A(2,I)= ',F7.2,'  DA(I)= ',F7.2)
 1455 CONTINUE
C
      DO 1460 I=SAITEMS,NITEMS
      WRITE(REPL,150) B(1,I),B(2,I)
  150 FORMAT(2F10.2)
```

```
1460 CONTINUE
     STOP
     END
```

APPENDIX D

PROGRAM 2

```
      PROGRAM AVERAGE(TAPE1,TAPE2,TAPE3,TAPE4)
C
C     ************************************
C     THIS PROGRAM AVERAGES THE SCALING
C     RESULTS ACROSS THREE REPLICATIONS
C     OF CHARACTERISTIC CURVE SCALING
C     ************************************
C
      REAL ALPHAI,BETAI,ALPHAT,BETAT
      REAL ABLT(3,2,1000),SUMAB(2),SIGMAB(2),AVGAB(2),SDAB(2)
      REAL UAVGB(3,2),UAVGA(3,2),UAVGC(3,2),AAVB(2),AAVA(2)
      REAL AAVC(2),AAVGB(3,2),AAVGA(3,2),AAVGC(3,2),TAVERB2(3)
      REAL AMINB(3,2),AMINA(3,2),AMAXB(3,2),AMAXA(3,2),UAVB(2)
      REAL UAVA(2),UAVC(2),B(3,2,2),A(3,2,2),AB(3,2,2),AA(3,2,2)
      REAL MMB(2,2),MMA(2,2),MMAB(2,2),MMAA(2,2),MMC,F(2),X(2)
      REAL UTRSC(3,2,1000),ATRSC(3,2,1000),USUMT(2),ASUMT(2)
      REAL USDT(2),UAVGT(2),ASDT(2),AAVGT(2),TUTRSCF(3,-50:50)
      REAL TUTRSCT(3,-50:50),SUMDIF,ASUMDIF,SUMDIF2,DIF,ADIF
      REAL DIF2,ALDIFF,BEDIFF,TUTRSCD(3,-50:50),Y1(3),Y2(3)
      REAL MDIF,MADIF,RMSDIF,CORRB(3),CORRA(3)
      REAL SUMSCD(-50:50),SUMSCT(-50:50),SUMSCF(-50:50)
      REAL DIFSE(-50:50),DIFPEE(-50:50)
      REAL TTRSCD(-50:50),TTRSCT(-50:50),TTRSCF(-50:50)
      REAL ABIL(-50:50),TE(-50:50)
      REAL ETRSC2(600),AABLT(600),B2(50),A2(50),DIFC(501)
      REAL ANCHB1(3,20),ANCHB2(3,20),AVERB1(20),AVERB2(20)
      REAL TTRSCG(600),TTRSCU(600),AAA(3,50),BBB(3,50),MTE
      REAL PE(501),PU(501),PG(501),CE(501),CU(501),CG(501)
      INTEGER NSUBJ,NSUBJ2,NSUBJ3,NITEMS,NAITEMS,REPL,AO
      INTEGER NUITEMS,NAITM2,NAITM3,AO2,AO3,HNSUBJ,TNSUBJ
      INTEGER TTNSUBJ,E(501),U(501),GG(501),SCORE(501),Z
C
      READ(1,5) NSUBJ,NITEMS,NAITEMS,REPL,AO1,ALPHAI,BETAI,
     1CORRB(1),CORRA(1)
    5 FORMAT(73(/),5I5/4F10.3)
C
      HNSUBJ=NSUBJ/2
      NUITEMS=NITEMS-NAITEMS
      TNSUBJ=HNSUBJ*3
      TTNSUBJ=TNSUBJ*2
      PRINT*,NUITEMS,HNSUBJ,TNSUBJ
      READ(2,10) NSUBJ2,NAITM2,AO2,ALPHAI,BETAI,CORRB(2),
     1CORRA(2)
   10 FORMAT(73(/),I5,5X,I5,5X,I5/4F10.3)
      READ(3,15) NSUBJ3,NAITM3,AO3,ALPHAI,BETAI,CORRB(3),
     1CORRA(3)
   15 FORMAT(73(/),I5,5X,I5,5X,I5/4F10.3)
      IF (NSUBJ.NE.NSUBJ2 .OR. NSUBJ.NE.NSUBJ3) PRINT*,'*'
      IF (NAITEMS.NE.NAITM2 .OR. NAITEMS.NE.NAITM3)
     1PRINT*,'**'
      IF (AO1.NE.AO2 .OR. AO1.NE.AO3) PRINT*,'***'
C
      DO 1000 N=1,3
```

```
      DO 1003 G=1,2
      READ(N,20) (ABLT(N,G,J),J=1,HNSUBJ)
   20 FORMAT(F12.3)
 1003 CONTINUE
 1000 CONTINUE
C
C     ...CALCULATE MEANS AND STANDARD DEVIATIONS
C        OF EXAMINEE ABILITY FOR EACH GROUP...
C
      DO 1004 G=1,2
      SUMAB(G)=0.0
      DO 1005 N=1,3
      DO 1010 J=1,HNSUBJ
      SUMAB(G)=SUMAB(G)+ABLT(N,G,J)
 1010 CONTINUE
 1005 CONTINUE
 1004 CONTINUE
      DO 1015 G=1,2
      AVGAB(G)=SUMAB(G)/FLOAT(TNSUBJ)
      SIGMAB(G)=0.0
      DO 1018 N=1,3
      DO 1020 J=1,HNSUBJ
      SIGMAB(G)=SIGMAB(G)+(ABLT(N,G,J)-AVGAB(G))**2
 1020 CONTINUE
 1018 CONTINUE
      SDAB(G)=SQRT(SIGMAB(G)/(FLOAT(TNSUBJ)-1.0))
      PRINT*,'SDAB(G)= ',SDAB(G)
 1015 CONTINUE
C
      DO 1025 N=1,3
      DO 1030 G=1,2
      READ(N,25) UAVGB(N,G),UAVGA(N,G),UAVGC(N,G),AAVGB(N,G),
     1AAVGA(N,G),AAVGC(N,G)
   25 FORMAT(6F6.3)
 1030 CONTINUE
 1025 CONTINUE
C
      DO 1035 N=1,3
      DO 1040 G=1,2
      READ(N,30) B(N,G,1),A(N,G,1),AB(N,G,1),AA(N,G,1),
     1B(N,G,2),A(N,G,2),AB(N,G,2),AA(N,G,2)
   30 FORMAT(8F10.3)
 1040 CONTINUE
 1035 CONTINUE
C
C     ...CALCULATE MEANS AND MIN/MAX FOR UNIQUE AND
C        ANCHOR ITEMS...
C
      DO 1050 G=1,2
      UAVB(G)=(UAVGB(1,G)+UAVGB(2,G)+UAVGB(3,G))/3.0
      UAVA(G)=(UAVGA(1,G)+UAVGA(2,G)+UAVGA(3,G))/3.0
      UAVC(G)=(UAVGC(1,G)+UAVGC(2,G)+UAVGC(3,G))/3.0
      AAVB(G)=(AAVGB(1,G)+AAVGB(2,G)+AAVGB(3,G))/3.0
      AAVA(G)=(AAVGA(1,G)+AAVGA(2,G)+AAVGA(3,G))/3.0
```

```
      AAVC(G)=(AAVGC(1,G)+AAVGC(2,G)+AAVGC(3,G))/3.0
C
      DO 1043 N=1,3
      DO 1045 M=1,2
      IF(B(N,G,M).EQ.99.0) THEN
      B(N,G,M)=0.0
      A(N,G,M)=0.8
      ENDIF
      IF(AB(N,G,M).EQ.99.0) THEN
      AB(N,G,M)=0.0
      AA(N,G,M)=0.8
      ENDIF
 1045 CONTINUE
 1043 CONTINUE
C
      IF (B(1,G,1).GT.B(2,G,1) .AND.
     1B(1,G,1).GT.B(3,G,1)) THEN
      MMB(1,G)=B(1,G,1)
      ELSE IF (B(2,G,1).GT.B(3,G,1) .AND.
     1B(2,G,1).GT.B(1,G,1)) THEN
      MMB(1,G)=B(2,G,1)
      ELSE
      MMB(1,G)=B(3,G,1)
      ENDIF
      IF (A(1,G,1).GT.A(2,G,1) .AND.
     1A(1,G,1).GT.A(3,G,1)) THEN
      MMA(1,G)=A(1,G,1)
      ELSE IF (A(2,G,1).GT.A(3,G,1) .AND.
     1A(2,G,1).GT.A(1,G,1)) THEN
      MMA(1,G)=A(2,G,1)
      ELSE
      MMA(1,G)=A(3,G,1)
      ENDIF
      IF (AB(1,G,1).GT.AB(2,G,1) .AND.
     1AB(1,G,1).GT.AB(3,G,1)) THEN
      MMAB(1,G)=AB(1,G,1)
      ELSE IF (AB(2,G,1).GT.AB(3,G,1) .AND.
     1AB(2,G,1).GT.AB(1,G,1)) THEN
      MMAB(1,G)=AB(2,G,1)
      ELSE
      MMAB(1,G)=AB(3,G,1)
      ENDIF
      IF (AA(1,G,1).GT.AA(2,G,1) .AND.
     1AA(1,G,1).GT.AA(3,G,1)) THEN
      MMAA(1,G)=AA(1,G,1)
      ELSE IF (AA(2,G,1).GT.AA(3,G,1) .AND.
     1AA(2,G,1).GT.AA(1,G,1)) THEN
      MMAA(1,G)=AA(2,G,1)
      ELSE
      MMAA(1,G)=AA(3,G,1)
      ENDIF
C
      IF (B(1,G,2).LT.B(2,G,2) .AND. B(1,G,2).LT.B(3,G,2)) THEN
      MMB(2,G)=B(1,G,2)
```

```
      ELSE IF (B(2,G,2).LT.B(3,G,2) .AND.
     1B(2,G,2).LT.B(1,G,2)) THEN
      MMB(2,G)=B(2,G,2)
      ELSE
      MMB(2,G)=B(3,G,2)
      ENDIF
      IF (A(1,G,2).LT.A(2,G,2) .AND.
     1A(1,G,2).LT.A(3,G,2)) THEN
      MMA(2,G)=A(1,G,2)
      ELSE IF (A(2,G,2).LT.A(3,G,2) .AND.
     1A(2,G,2).LT.A(1,G,2)) THEN
      MMA(2,G)=A(2,G,2)
      ELSE
      MMA(2,G)=A(3,G,2)
      ENDIF
      IF (AB(1,G,2).LT.AB(2,G,2) .AND.
     1AB(1,G,2).LT.AB(3,G,2)) THEN
      MMAB(2,G)=AB(1,G,2)
      ELSE IF (AB(2,G,2).LT.AB(3,G,2) .AND.
     1AB(2,G,2).LT.AB(1,G,2)) THEN
      MMAB(2,G)=AB(2,G,2)
      ELSE
      MMAB(2,G)=AB(3,G,2)
      ENDIF
      IF (AA(1,G,2).LT.AA(2,G,2) .AND.
     1AA(1,G,2).LT.AA(3,G,2)) THEN
      MMAA(2,G)=AA(1,G,2)
      ELSE IF (AA(2,G,2).LT.AA(3,G,2) .AND.
     1AA(2,G,2).LT.AA(1,G,2)) THEN
      MMAA(2,G)=AA(2,G,2)
      ELSE
      MMAA(2,G)=AA(3,G,2)
      ENDIF
      MMC=0.2
 1050 CONTINUE
C
      DO 1060 N=1,3
      READ(N,35) Y1(N),Y2(N),ALPHAT,BETAT
   35 FORMAT(2F10.3/2F10.3)
      SUMX1=SUMX1+Y1(N)
      SUMX2=SUMX2+Y2(N)
 1060 CONTINUE
      X(1)=SUMX1/3.0
      X(2)=SUMX2/3.0
      ALDIFF=ALPHAT-X(1)
      BEDIFF=BETAT-X(2)
      DO 1065 G=1,2
      DO 1068 N=1,3
      DO 1070 J=1,HNSUBJ
      READ(N,40) UTRSC(N,G,J)
   40 FORMAT(F6.3)
 1070 CONTINUE
 1068 CONTINUE
 1065 CONTINUE
```

```
      DO 1075 G=1,2
      DO 1076 N=1,3
      DO 1078 J=1,HNSUBJ
      READ(N,45) ATRSC(N,G,J)
   45 FORMAT(F6.3)
 1078 CONTINUE
 1076 CONTINUE
 1075 CONTINUE
C
C     ...CALCULATE MEANS AND STANDARD DEVIATIONS OF
C        TRUE SCORES ON UNIQUE ITEMS...
C
      DO 1080 G=1,2
      USUMT(G)=0.0
      DO 1082 N=1,3
      DO 1085 J=1,HNSUBJ
      USUMT(G)=USUMT(G)+UTRSC(N,G,J)
 1085 CONTINUE
 1082 CONTINUE
 1080 CONTINUE
      DO 1090 G=1,2
      USIGT=0.0
      USDT(G)=0.0
      UAVGT(G)=USUMT(G)/FLOAT(TNSUBJ)
      DO 1093 N=1,3
      DO 1095 J=1,HNSUBJ
      USIGT=USIGT+(UTRSC(N,G,J)-UAVGT(G))**2
 1095 CONTINUE
 1093 CONTINUE
      USDT(G)=SQRT(USIGT/(FLOAT(TNSUBJ)-1.0))
 1090 CONTINUE
C
C     ...CALCULATE MEANS AND STANDARD DEVIATIONS OF
C        TRUE SCORES ON ANCHOR ITEMS...
C
      DO 1100 G=1,2
      ASUMT(G)=0.0
      DO 1103 N=1,3
      DO 1105 J=1,HNSUBJ
      ASUMT(G)=ASUMT(G)+ATRSC(N,G,J)
 1105 CONTINUE
 1103 CONTINUE
 1100 CONTINUE
      DO 1110 G=1,2
      ASIGT=0.0
      AAVGT(G)=ASUMT(G)/FLOAT(TNSUBJ)
      DO 1112 N=1,3
      DO 1115 J=1,HNSUBJ
      ASIGT=ASIGT+(ATRSC(N,G,J)-AAVGT(G))**2
 1115 CONTINUE
 1112 CONTINUE
      ASDT(G)=SQRT(ASIGT/(FLOAT(TNSUBJ)-1.0))
 1110 CONTINUE
C
```

```
      AVCORRB=(CORRB(1)+CORRB(2)+CORRB(3))/3.0
      AVCORRA=(CORRA(1)+CORRA(2)+CORRA(3))/3.0
C
      WRITE(4,60)
   60 FORMAT(13X,50('-')//25X,'AVERAGE SCALING RESULTS'//
     113X,50('-'))
      WRITE(4,65) HNSUBJ,NITEMS,NAITEMS,AO
   65 FORMAT(///15X,'SAMPLE SIZE / GROUP =',I5/15X,'TOTAL
     1NUMBER OF ITEMS =',I5/15X,'NUMBER OF ANCHOR ITEMS ='
     2,I5/15X,'ABILITY OVERLAP =',I5,'%')
      WRITE(4,70) X(1),X(2),ALPHAT,BETAT,ALDIFF,BEDIFF
   70 FORMAT(//15X,'CALCULATED ALPHA=',F6.2,5X,'CALCULATED
     1BETA=',F6.2///15X,'TRUE ALPHA=',F6.2,5X,'TRUE BETA=',
     2F6.2///15X,'TRUE-CALCULATED ALPHA=',F6.2,5X,
     3'TRUE-CALCULATED BETA=',F6.2)
      WRITE(4,75) TNSUBJ,AVGAB(1),SDAB(1),AVGAB(2),SDAB(2)
   75 FORMAT(////14X,'SUMMARY STATISTICS OF EXAMINEE ABILITY
     1(N=',I5,')'//13X,50('-')/16X,'GROUP',14X,'MEAN',
     214X,'SD'/13X,50('-')//18X,'1',15X,F4.2,14X,F4.2/
     218X,'2',15X,F4.2,14X,F4.2//13X,50('-'))
      WRITE(4,85) NUITEMS,UAVB(1),MMB(1,1),UAVA(1),MMA(1,1),
     +UAVC(1),MMC,MMB(2,1),MMA(2,1),MMC,UAVB(2),MMB(1,2),
     +UAVA(2),MMA(1,2),UAVC(2),MMC,MMB(2,2),MMA(2,2),MMC
   85 FORMAT(//9X,'SUMMARY STATISTICS OF UNIQUE ITEM
     1PARAMETERS(N=',I5,')'/5X,68('-')/27X,'B',17X,'A',
     217X,'C'/5X,68('-')/8X,'GROUP',8X,'MEAN',5X,'MAX/
     2',5X,'MEAN',5X,'MAX/',5X,'MEAN',5X,'MAX/'/
     330X,'MIN',15X,'MIN',15X,'MIN'/5X,68('-')//
     410X,'1',9X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2,5X,F5.2,3X,
     5F5.2/28X,F5.2,13X,F5.2,13X,F5.2//
     610X,'2',9X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2,5X,F5.2,3X,
     7F5.2/28X,F5.2,13X,F5.2,13X,F5.2//
     85X,68('-'))
      WRITE(4,90) NAITEMS,AAVB(1),MMAB(1,1),AAVA(1),MMAA(1,1),
     +AAVC(1),MMC,MMAB(2,1),MMAA(2,1),MMC,AAVB(2),MMAB(1,2),
     +AAVA(2),MMAA(1,2),AAVC(2),MMC,MMAB(2,2),MMAA(2,2),MMC
   90 FORMAT(//9X,'SUMMARY STATISTICS OF ANCHOR ITEM
     1PARAMETERS(N=',I5,')'/5X,68('-')/27X,'B',17X,'A',17X,
     2'C'/5X,68('-')/8X,'GROUP',8X,'MEAN',5X,'MAX/',5X,'MEAN',
     3X,'MAX/',5X,'MEAN',5X,'MAX/'/
     430X,'MIN',15X,'MIN',15X,'MIN'/5X,68('-')//
     510X,'1',9X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2/
     628X,F5.2,13X,F5.2,13X,F5.2//
     710X,'2',9X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2,5X,F5.2,3X,F5.2/
     828X,F5.2,13X,F5.2,13X,F5.2//
     95X,68('-'))
      WRITE(4,95) UAVGT(1),USDT(1),AAVGT(1),ASDT(1),UAVGT(2),
     1USDT(2),AAVGT(2),ASDT(2)
   95 FORMAT(//9X,'SUMMARY STATISTICS OF TRUE SCORES'/
     15X,68('-')/25X,'UNIQUE ITEMS',12X,'ANCHOR ITEMS'/
     25X,68('-')/10X,'GROUP',11X,'MEAN',3X,'SD',15X,
     3'MEAN',3X,'SD'/5X,68('-')//12X,'1',11X,2F6.2,12X,2F6.2
     4//12X,'2',11X,2F6.2,12X,2F6.2//
     55X,68('-'))
```

```
      WRITE(4,98)AVCORRB,AVCORRA
   98 FORMAT(//9X,'CORRELATIONS OF ANCHOR ITEM PARAMETERS
     1FOR'/9X,'TESTS 1 AND 2'/5X,68('-')//
     210X,'PARAMETER',20X,'CORRELATION'/
     35X,68('-')//14X,'B',20X,F10.3//14X,'A',20X,F10.3/
     45X,68('-'))
C
      BIT=1.0
      DO 1140 N=1,3
      DO 1145 ABL=-30, 30, BIT
      READ(N,99)ABIL(ABL),TUTRSCD(N,ABL),
     1TUTRSCT(N,ABL),TUTRSCF(N,ABL)
   99 FORMAT(4F10.3)
 1145 CONTINUE
 1140 CONTINUE
      DO 1150 ABL=-30, 30, BIT
      SUMSCD(ABL)=0.0
      SUMSCT(ABL)=0.0
      SUMSCF(ABL)=0.0
      DO 1155 N=1,3
      SUMSCD(ABL)=SUMSCD(ABL)+TUTRSCD(N,ABL)
      SUMSCT(ABL)=SUMSCT(ABL)+TUTRSCT(N,ABL)
      SUMSCF(ABL)=SUMSCF(ABL)+TUTRSCF(N,ABL)
 1155 CONTINUE
      TTRSCD(ABL)=SUMSCD(ABL)/3.0
      TTRSCT(ABL)=SUMSCT(ABL)/3.0
      TTRSCF(ABL)=SUMSCF(ABL)/3.0
 1150 CONTINUE
C
C     ...CALCULATE INDICES OF SCALING ERROR...
C
      SUMDIF=0.0
      ASUMDIF=0.0
      SUMDIF2=0.0
      DO 1160 ABL=-30, 30, BIT
      DIFSE(ABL)=TTRSCT(ABL)-TTRSCF(ABL)
      SUMDIF=SUMDIF+DIFSE(ABL)
      ADIF=ABS(TTRSCT(ABL)-TTRSCF(ABL))
      ASUMDIF=ASUMDIF+ADIF
      DIF2=(TTRSCT(ABL)-TTRSCF(ABL))**2
      SUMDIF2=SUMDIF2+DIF2
 1160 CONTINUE
      MDIF=SUMDIF/61.0
      AMDIF=ASUMDIF/61.0
      RMSDIF1=SQRT(SUMDIF2/61.0)
C
      WRITE(4,120) MDIF,AMDIF,RMSDIF1
  120 FORMAT(///9X,'SCALING ERROR INDICES :(TTRSCT-
     1TTRSCF)'/5X,68('-')//25X,'MEAN DIFFERENCE= ',F6.2,
     2/25X,'MEAN ABSOLUTE DIFFERENCE= ',F6.2,25X,
     3/'ROOT MEAN SQUARE DIFFERENCE= ',F6.2//5X,68('-'))
C
C     ...CALCULATE INDICES OF PARAMETER ESTIMATION ERROR...
C
```

```
      SUMDIF=0.0
      ASUMDIF=0.0
      SUMDIF2=0.0
      DO 1170 ABL=-30, 30, BIT
      DIFPEE(ABL)=TTRSCD(ABL)-TTRSCT(ABL)
      SUMDIF=SUMDIF+DIFPEE(ABL)
      ADIF=ABS(TTRSCD(ABL)-TTRSCT(ABL))
      ASUMDIF=ASUMDIF+ADIF
      DIF2=(TTRSCD(ABL)-TTRSCT(ABL))**2
      SUMDIF2=SUMDIF2+DIF2
 1170 CONTINUE
      MDIF=SUMDIF/61.0
      AMDIF=ASUMDIF/61.0
      RMSDIF2=SQRT(SUMDIF2/61.0)
C
      WRITE(4,130) MDIF,AMDIF,RMSDIF2
  130 FORMAT(///9X,'PARAMETER EST ERROR INDICES :(TTRSCD-
     1TTRSCT)'/5X,68('-')//25X,'MEAN DIFFERENCE= ',F6.2,
     2/25X,'MEAN ABSOLUTE DIFFERENCE= ',F6.2,25X,
     3/'ROOT MEAN SQUARE DIFFERENCE= ',F6.2//5X,68('-'))
C
C     ...CALCULATE INDICES OF TOTAL ERROR...
C
      DO 1171 ABL=-30,30,BIT
      ADIF1=ABS(TTRSCT(ABL)-TTRSCF(ABL))
      ADIF2=ABS(TTRSCD(ABL)-TTRSCT(ABL))
      TE(ABL)=ADIF1+ADIF2
      SUMTE=SUMTE+TE(ABL)
 1171 CONTINUE
      RMTE=RMSDIF1+RMSDIF2
      MTE=SUMTE/61.0
      WRITE(4,132) MTE,RMTE
  132 FORMAT(///9X,'TOTAL ERROR INDICES :MAD(SE)+MAD(PEE)',
     1' & RMSTE(SE)+RMSTE(PEE)'/5X,68('-')//
     225X,'MEAN TOTAL ERROR = ',F6.2//
     325X,'RMSQ TOTAL ERROR = ',F6.2//5X,68('-'))
C
      WRITE(4,135)
  135 FORMAT(///,7X,'ABL',5X,'TTRSCD',5X,'TTRSCT',5X,'TTRSCF',10X,
     1'PEE',6X,'SE',7X,'TE'/7X,'---',5X,'------',5X,
     2'------',5X,'------',10X,'---',6X,'--',7X,'--'/)
      DO 1172 N=4,6,2
      DO 1175 ABL=-30, 30, BIT
      WRITE(N,138)ABIL(ABL),TTRSCD(ABL),TTRSCT(ABL),TTRSCF(ABL),
     1DIFPEE(ABL),DIFSE(ABL),TE(ABL)
  138 FORMAT(7X,F4.1,6X,F4.1,7X,F4.1,6X,F4.1,9X,F5.2,4X,F5.2,
     14X,F5.2)
 1175 CONTINUE
 1172 CONTINUE
C
C     ...CALCULATE PERCENTILE RANKS OF TRANSFORMED
C        TRUE SCORES FOR GROUP 2...
C
      G=2
```

```
      DO 1180 J=1,HNSUBJ
      SUTRSC=0.0
      SABLT=0.0
      DO 1185 N=1,3
      SUTRSC=SUTRSC+UTRSC(N,G,J)
      SABLT=SABLT+ABLT(N,G,J)
 1185 CONTINUE
      ETRSC2(J)=SUTRSC/3.0
      AABLT(J)=SABLT/3.0
      ETRSC2(J)=INT((ETRSC2(J)+0.05)*10.0)
 1180 CONTINUE
C
      DO 1186 N=1,3
      DO 1188 I=1,NUITEMS
      READ(N,140) BBB(N,I),AAA(N,I)
  140 FORMAT(2F10.2)
 1188 CONTINUE
 1186 CONTINUE
C
      DO 1190 I=1,NUITEMS
      SUB=0.0
      SUA=0.0
      DO 1195 N=1,3
      IF(BBB(N,I).EQ.99.0) GOTO 1195
      SUB=SUB+BBB(N,I)
      SUA=SUA+AAA(N,I)
 1195 CONTINUE
      B2(I)=SUB/3.0
      A2(I)=SUA/3.0
 1190 CONTINUE
C
      DO 1200 J=1,HNSUBJ
      TTRSCG(J)=0.0
      TTRSCU(J)=0.0
      TABT=AABLT(J)*ALPHAT+BETAT
      DO 1205 I=1,NUITEMS
      ATRANF=A2(I)/X(1)
      BTRANF=B2(I)*X(1)+X(2)
      POW=1.7*(TABT-BTRANF)*ATRANF
      PR=0.2+0.8*EXP(POW)/(1.0+EXP(POW))
      TTRSCG(J)=TTRSCG(J)+PR
C
      ATRANT=A2(I)/ALPHAT
      BTRANT=B2(I)*ALPHAT+BETAT
      POW=1.7*(TABT-BTRANT)*ATRANT
      PR=0.2+0.8*EXP(POW)/(1.0+EXP(POW))
      TTRSCU(J)=TTRSCU(J)+PR
 1205 CONTINUE
      TTRSCG(J)=INT((TTRSCG(J)+0.05)*10.0)
      TTRSCU(J)=INT((TTRSCU(J)+0.05)*10.0)
 1200 CONTINUE
C
      DO 1210 K=0,500
      E(K)=0
```

```
      U(K)=0
      GG(K)=0
      PE(K)=0.0
      PU(K)=0.0
      PG(K)=0.0
 1210 CONTINUE
      DO 1212 P=0,500
      CE(P)=0.0
      CU(P)=0.0
      CG(P)=0.0
 1212 CONTINUE
      DO 1215 J=1,HNSUBJ
      DO 1220 K=0,500
      IF(ETRSC2(J).EQ.K)  E(K)=E(K)+1
      IF(TTRSCU(J).EQ.K)  U(K)=U(K)+1
      IF(TTRSCG(J).EQ.K)  GG(K)=GG(K)+1
 1220 CONTINUE
 1215 CONTINUE
      DO 1225 K=0,500
      PE(K)=(E(K)/300.0)*100.0
      PU(K)=(U(K)/300.0)*100.0
      PG(K)=(GG(K)/300.0)*100.0
 1225 CONTINUE
      DO 1227 P=0,500
      DO 1228 K=0,P
      CE(P)=CE(P)+PE(K)
      CU(P)=CU(P)+PU(K)
      CG(P)=CG(P)+PG(K)
 1228 CONTINUE
 1227 CONTINUE
      DO 1230 P=1,500
      Z=P/10
      SCORE(P)=Z
 1230 CONTINUE
C
      DO 1235 P=0,500,10
      DIFC(P)=CU(P)-CG(P)
 1235 CONTINUE
      WRITE(4,145)
  145 FORMAT(//18X,'TTRSCT',16X,'TTRSCF'/
     118X,6('-'),16X,6('-')/
     27X,'SCORE',8X,'CF',
     320X,'CF',11X,'DIFC'/5X,68('-'))
      DO 1240 P=50,500,50
      WRITE(4,150)SCORE(P),CU(P),CG(P),DIFC(P)
  150 FORMAT(7X,I3,6X,F6.1,16X,F6.1,9X,F6.1)
 1240 CONTINUE
      WRITE(4,152)
  152 FORMAT(5X,68('-'))
      WRITE(4,155)
  155 FORMAT(//18X,'TTRSCT',16X,'TTRSCF'/
     118X,6('-'),16X,6('-')/
     27X,'SCORE',8X,'CF',
     320X,'CF',11X,'DIFC'/5X,68('-'))
```

```
      DO 1250 P=0,500,10
      WRITE(4,160)SCORE(P),CU(P),CG(P),DIFC(P)
  160 FORMAT(7X,I3,6X,F6.1,16X,F6.1,9X,F6.1)
 1250 CONTINUE
      WRITE(4,162)
  162 FORMAT(5X,68('-'))
C
      DO 1255 I=1,NAITEMS
      SUMB1=0.0
      SUMB2=0.0
      NCNT=3
      DO 1260 N=1,3
      READ(N,165) ANCHB1(N,I),ANCHB2(N,I)
  165 FORMAT(2F10.3)
      IF(ANCHB1(N,I).EQ.99.0) THEN
      NCNT=NCNT-1
      GOTO 1260
      ELSE
      SUMB1=SUMB1+ANCHB1(N,I)
      SUMB2=SUMB2+ANCHB2(N,I)
      END IF
 1260 CONTINUE
      AVERB1(I)=SUMB1/FLOAT(NCNT)
      AVERB2(I)=SUMB2/FLOAT(NCNT)
      TAVERB2(I)=AVERB2(I)*ALPHAT+BETAT
 1255 CONTINUE
      WRITE(4,170)
  170 FORMAT(///,9X,'ANCHOR ITEM DIFFICULTY VALUES'/
     15X,68('-'), /18X,'GROUP1',10X,'GROUP2'/18X,
     2'------',10X,'------',/)
      DO 1265 I=1,NAITEMS
      WRITE(4,180) AVERB1(I),AVERB2(I),TAVERB2(I)
  180 FORMAT(17X,F6.2,6X,F6.2,2X,F6.2)
 1265 CONTINUE
      WRITE(4,185)
  185 FORMAT(5X,68('-'))
C
      RETURN
      END
```

REFERENCES

Anastasi, A. (1954). Psychological testing. New York, NY: Macmillan Company.

Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. (Report No. 88-2). New York, NY: College Entrance Examination Board.

Bejar, I. I. (1988). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 22, 21-31.

Beller, M. (1990, April . Psychometric issues in admission procedures of Israeli universities. Paper presented at the meeting of AERA, Boston.

Berry, G. L., & Lopez, C. A. (1977). Testing programs and the Spanish-speaking child: Assessment guidelines for school counselors. The School Counselor, March, 261-269.

Blanton, R. L. (1975). Historical perspective on classification of mental retardation. In N. Hobbs (Ed.), Issues in the Classification of Children. San Francisco, CA: Jossey-Bass.

Candell, G. L., & Hulin, C. L. (1986). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. Journal of Cross-Cultural Psychology, 17(4), 417-440.

Cook, L. L., & Eignor, D. R. (1983, April). An investigation of the feasibility of applying item response theory to equate achievement tests. Paper presented at the meeting of AERA, Montreal.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 175-195). Vancouver, BC: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. International Journal of Educational Research, 13, 161-173.

Cook, L. L., Eignor, D. R., & Petersen, N. S. (1985). A study of the temporal stability of item parameter estimates (ETS Research Report 85-45). Princeton, NJ: Educational Testing Service.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: CBS College Publishing.

Darcy, N. T. (1953). A review of the literature on the effects of bilingualism upon the measure of intelligence. Journal of Genetic Psychology, 82, 21-57.

Darcy, N. T. (1963). Bilingualism and the measure of intelligence: Review of a decade of research. Journal of Genetic Psychology, 103, 259-282.

DeAvila, E. A., & Havassy, B. (1974). The testing of minority children - a neo-Piagetian approach. Today's Education, December, 72-75.

DeBlassie, R. R. (1980). Testing Mexican American youth. Hingham, MA: Teaching Resources Corporation.

Diaz, R. M. (1983). Thought and two languages: The impact of bilingualism on cognitive development. In E. W. Gordon (Ed.), Review of research in education, 10. Washington, DC: American Educational Research Association.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. Applied Psychological Measurement, 9, 413-415.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74, 912-921.

Fletcher, R., & Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. The Computer Journal, 6, 163-168.

Fouad, A. N., & Hansen, J. C. (1987). Cross-cultural predictive accuracy of the Strong Campbell Interest Inventory. Measurement and Evaluation in Counseling and Development, 20, 3-10.

Grassau, E. (1969). Educational Testing in Chile. In K. Ingenkamp (Ed.), Developments in educational testing (volume 1). New York, NY: Science Publishers.

Gross, L. J., & Scott, J. W. (1989). Translating a health professional certification test to another language: A pilot analysis. Evaluation and the Health Professions, 12, 61-72.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade equivalent scale for the vertical equating of test scores. Applied Psychological Measurement, 5, 187-201.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Murray, L. L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 71-94). Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of the IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

Hambleton, R. K., & Rovinelli, R. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavioral Science, 17, 73-74.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 48, 467-510.

Hansen, J. J. (1987). Cross-cultural research on vocational interests. Measurement and Evaluation in Counseling and Development, 19, 163-176.

Hansen, J. C., & Fouad, A. N. (1984). Translation and validation of the Spanish form of the Strong-Campbell Interest Inventory. Measurement and Evaluation in Counseling and Development, 16, 192-197.

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Hulin, C. L. (1987). A psychometric theory of evaluation of item and scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology, 18, 115-142.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. Journal of Applied Psychology, 67, 818-825.

Hulin, C. L., & Mayer, L. M. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. Journal of Applied Psychology, 71(1), 84-94.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed)., Applications of item response theory (pp. 155-174). Vancouver, BC: Educational Research Institute of British Columbia.

Irvine, S. H., & Carroll, W. K. (1980). Testing and assessment across cultures: Issues in methodology and theory. In H. Triandis & J. Berry (Eds.), Handbook of cross-cultural psychology, volume 2: Methodology. Boston, MA: Allyn & Bacon.

Johansen, G. A. (1987). A study of item response theory equating with an anchor test design. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Jones, D. M. (1989). Culture and testing. American Psychologist, February, 360-366.

Katerburg, R., Hoy, S., & Smith, F. J. (1977). Language, time and person effects on attitude scale translations. Journal of Applied Psychology, 71(4), 385-391.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. Journal of Educational Measurement, 3, 22, 197-206.

Kline, P. (1983). The cross-cultural use of personality tests. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors. New York, NY: Plenum Publishers.

Kolen, M. J. (1981). Comparisons of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 18, 1-11.

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the Tests of General Educational Development. Journal of Educational Measurement, 19, 279-293.

Kulkarni, S. S. (1969). Constructing equivalent achievement survey tests to be used for different student populations in India. In K. Ingenkamp (Ed.), Developments in educational testing (Volume 1). New York, NY: Science Publishers, Inc.

Lapointe, A. E., Mead, N. A., & Phillips, G. W. (1989). A world of differences: An international assessment of mathematics and science (Report No. 19-CAEP-01). Princeton, NJ: Educational Testing Service.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, T. L. (1981). An investigation of item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Lord, F. M. (1977). Practical applications of item response theory. Journal of Educational Measurement, 14, 117-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." Applied Psychological Measurement, 8, 453-461.

Marco, G. L., Peterson, N. S., & Stewart, E. E. (1979, April). Applicability of two logistic models for equating test scores when tests and samples are varied. Paper presented at the meeting of AERA, San Francisco.

Mayberry, P. W. (1984, April). Analysis of cross-cultural attitudinal scale translation using maximum likelihood factor analysis. Paper presented at the meeting of AERA, New Orleans.

McCauley, D. E., & Colberg, M. (1983). Transportability of deductive measurement across cultures. Journal of Educational Measurement, 20, 81-92.

McDonald, R. P. (1980). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 33, 205-233.

McKinley, R. L., & Reckase, M. D. (1981). A comparison of procedures for constructing large item pools (Research Report 81-3). Columbia, MO: University of Missouri, Department of Educational Psychology.

Mellenbergh, G. J. (1983). Conditional item bias methods. In S. H. Irving & J. W. Berry (Eds.), Human assessment and cultural factors. New York, NY: Plenum Publishers.

Mellenbergh, G. J. (1989). Item bias and item response theory. International Journal of Educational Research, 13, 127-143.

Mercer, J. R. (1979). SOMPA technical manual. New York, NY: Psychological Corporation.

Mitchell, J. V. (Ed.) (1985). Ninth mental measurement yearbook. Lincoln, NE: University of Nebraska Press.

Olmedo, E. L. (1981). Testing linguistic minorities. American Psychologist, 36, 1078-1085.

Padilla, A. M. (1979). Critical factors in the testing of Hispanic Americans: A review and some suggestions for the future. In R. W. Tyler & S. H. White (Eds.), Testing, teaching, and learning: Report of a conference on testing. Washington, DC: National Institute of Education.

Perez, F. M. (1980). Performance of bilingual children on the Spanish version of the ITPA. Exceptional Children, 46, 536-541.

Peterson, N. S., Cook, L. L., & Stocking, M. S. (1981, April). Scale drift: A comparative study of IRT versus linear equating methods. Paper presented at the meeting of AERA, Los Angeles.

Phillips, S. E. (1985). Quantifying equating errors with item response methods. Applied Psychological Measurement, 9, 303-317.

Poortinga, Y. H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors. New York, NY: Plenum Publishers.

Quay, L. (1971). Language, dialect reinforcement and the intelligence test performance of Negro children. Child Development, 42, 5-15.

Raju, N. S., Edwards, J. E., Osberg, D. W. (1983, April). The effect of anchor test size in vertical equating with the Rasch and three-parameter models. Paper presented at the meeting of NCME, Montreal.

Raphael, P. (1989, March). Ontario IAEP results: Comparable data for English and French speaking students. Paper presented at the meeting of AERA, San Francisco.

Roca, P. (1955). Problems of adapting intelligence scales from one culture to another. High School Journal, 38, 124-131.

Rogers, H. J., & Hambleton, R.K. (1989). Evaluation of computer simulated baseline statistics for use in item bias studies. Educational and Psychological Measurement, 49, 355-369.

Samuda, R. J. (1975). Psychological testing of American minorities: Issues and consequences. New York, NY: Dodd, Mead & Company.

Samuda, R. J. (1983). Cross-cultural testing within a multicultural society. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors. New York, NY: Plenum Publishers.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criterion. Journal of Educational Statistics, 6, 317-376.

Scheunemann, J. (1979). A new method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Simon, M. G. (1989, March). Bias in translated items: A comparative study of five statistical methods. Paper presented at the meeting of AERA, San Francisco.

Skaggs, G., & Lissitz, R. W. (1986). Test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Stanley, J. C., & Hopkins, K. D. (1972). Educational and psychological measurement and evaluation. Englewood Cliffs, NJ: Prentice Hall.

Swanson, H. L., & Watson, B. L. (1982). Educational and psychological assessment of exceptional children. London: C. V. Mosby Company.

Tamayo, J. M. (1987). Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation. Educational and Psychological Measurement, 47, 893-902.

Trimble, J. E., Lonner, W. J., & Boucher, J. D. (1983). Stalking the wily emic: Alternatives to cross-cultural measurement. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cultural factors. New York, NY: Plenum Publishers.

Vale, C. D., Marcelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. (1981). Methods for linking item parameters (AFHRL-TR-81-10). Brooks Air Force Base, TX: Air Force Human Resource Laboratory.

van de Flier, H. (1982). Deviant response patterns and comparability of test scores. Journal of Cross-Cultural Psychology, 13, 267-298.

van de Vijver, F. J. R., & Poortinga, Y. H. (1985). A comment on McCauley and Colberg's conception of cross-cultural transportability of tests. Journal of Educational Measurement, 22, 157-167.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), Advances in educational and psychological testing. Boston, MA: Kluwer Academic Publishers.

Weiner, F. D., Lernau, L. E., & Ernay, E. (1983). Measuring language competency in speakers of Black American English. Journal of Speech and Hearing Disorders, 48, 76-84.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8, 347-364.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). Specifying the characteristics of linking items used for item response theory item calibration (ETS Research Report 87-24). Princeton, NJ: Educational Testing Service.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.

Zirkel, P. R. (1972). Spanish-speaking students and standardized tests. The Urban Review, June, 32-40.